Title of Document:	**Buttressing a New Paradigm in Protein Folding: Experimental Tools to Distinguish between Downhill and Multi-State Folding Mechanisms**
	Nagalakshmi, Tiruvarur Sooriyanarayanan, Ph.D., 2014
Directed By:	Professor, Victor Muñoz National Center for Biotechnology - Spanish National Research Council (CNB-CSIC) & CEI UAM+CSIC
	Dr. Mourad Sadqi Scientist, National Center for Biotechnology - Spanish National Research Council (CNB-CSIC)

ABSTRACT

Many single-domain proteins fold in milliseconds or longer. However, the advent of fast folding kinetic techniques has permitted to identify many other proteins that fold in the order of (few) microseconds and thus very closely to the folding speed limit. This suggests that the proteins that fold in microsecond timescale either cross a marginal single free energy barrier, multiple very small barriers (multi-state), or no barrier at all (downhill). This results in the potential observation of broad complex unfolding transitions in these ultrafast folding proteins (in contrast to simple two-state behavior). Many of the ultrafast folding proteins have small size and fold into simple alpha helix-bundle topologies. Theoretical studies support the size scaling of protein folding barriers. Engrailed homeodomain, a 61-residue α-helical domain with a helix-turn-helix topology folds in microseconds and exhibits an apparently complex (un)folding process. The observed complexity in the (un)folding behavior of engrailed homeodomain rules out a simple two-state model, but the folding mechanism of this protein has been interpreted with a conventional three-state model.

The current work aims to develop a set of experimental and analytical methods that can determine unambiguously whether an apparently complex folding process of a fast folding protein is downhill or multi-state using engrailed homeodomain as a model. A large-scale multiple probe approach that combines equilibrium, fast-folding measurement and single molecule measurements has been used to provide critical information to unravel the mechanistic details of the folding mechanism of this protein. Double perturbation measurement on engrailed, in which the protein was unfolded by both chemical denaturant and temperature, showed

complex results. Multi-probe equilibrium thermal and chemical unfolding measurements on engrailed revealed differences in the melting temperature and chemical denaturation midpoints respectively. All these signatures conformed to downhill folding mechanism or existence of low-barrier(s). The estimated overall barrier height was ~ 0.5 RT near T_m, by globally fitting the entire equilibrium thermal unfolding data to Mean Field Model. Multi-probe temperature jump studies resulted in single exponential relaxations by infrared and non-exponential relaxations by fluorescence and probe-dependent kinetic amplitudes for the slow rates. This result could still be explained by a downhill behavior by globally fitting both the equilibrium and the kinetic data using the same model. Single molecule FRET measurements explored the transition path of engrailed near C_m and further confirmed the existence of downhill behavior with the estimated marginal barrier of < 1 RT. These results emphasize the importance of multi-probe measurements and appropriate utilization of statistical mechanical for analysis for fast-folding proteins.

RESUMEN

Gran variedad de proteínas monodominio se pliegan a partir de la escala de los milisegundos, sin embargo, la aparición de técnicas cinéticas capaces de analizar el plegamiento rápido ha permitido la identificación de muchas otras que pliegan en el orden de (pocos) microsegundos y, por tanto, muy cerca de la velocidad límite de plegamiento. Esto sugiere que las proteínas que pliegan en la escala de los microsegundos o bien atraviesan una pequeña barrera marginal de energía libre, o bien múltiples y pequeñas barreras (multi-estado), o bien no atraviesan del todo ninguna barrera (*downhill*). Estas posibilidades resultan en la observación de un amplio y complejo rango de transiciones de desplegamiento en dichas proteínas, en claro contraste con el comportamiento simple esperado para procesos de tipo dos estados. Un gran número de proteínas con plegamiento ultrarrápido se caracterizan a su vez por poseer un pequeño tamaño y una estructura nativa organizada en topologías simples consistentes en paquetes de α-hélices (*α-helix-bundle*). Por otra parte, estudios teóricos sostienen que existe una relación directa entre el tamaño y las barreras del plegamiento de las proteínas.

El homeodominio *engrailed*, proteína α-helicoidal de 61 residuos con topología hélice-giro-hélice, es capaz de plegarse en microsegundos y en experimentos realizados por otros grupos de investigación presenta un complejo proceso de (des)plegamiento. La complejidad observada en el comportamiento del (des)plegamiento del dominio ha llevado a interpretar el mismo como un modelo convencional tres estados como extensión directa del modelo simple dos estados con el que se le caracterizó originalmente.

El trabajo actual pretende implementar un conjunto de métodos experimentales y analíticos de alta resolución que puedan determinar inequívocamente si un proceso de plegamiento de una proteína con plegamiento rápido aparentemente complejo es *downhill* o multi-estado utilizando el homeodominio *engrailed* como modelo. Para desenmarañar los detalles del mecanismo de plegamiento de esta proteína se ha obtenido información a partir de un abordaje experimental que utiliza multiples sondas espectroscópicas combinadas en medidas en equilibrio, medidas de plegamiento rápido y medidas de fluorescencia en molécula única. Por un lado, las medidas de perturbación doble en equilibrio realizadas sobre el homeodominio *engrailed*, en las cuales la proteína fue desplegada mediante desnaturalizantes químicos y mediante temperatura, mostraron resultados complejos compatibles con procesos de plegamiento tipo *downhill*. Asimismo, el análisis en equilibrio de desnaturalización térmica utilizando sondas espectróscopicas múltiples, así como del desplegamiento químico sobre el dominio, revelaron diferencias en T_m y C_m, respectivamente, los cuales son nuevamente indicios de plegamiento tipo *downhill*. También se han realizado estudios cinéticos utilizando el método de salto de temperatura en nanosegundos y midiendo el proceso mediante técnicas de espectroscopía de infrarrojos y fluorescencia. Los experimentos utilizando infrarrojos resultaron en relajaciones exponenciales simples mientras que los de fluorescencia mostraron relajaciones no exponenciales ajustables a decaimientos exponenciales dobles. Los estudios mostraron que la amplitud cinética relativa de la fase más lenta cambiaba significativamente en función de la sonda utilizada. Los resultados cinéticos pueden ser explicados mediante un comportamiento *downhill*,

atribuyendo la fase más rápida a la presencia de una barrera marginal. El análisis global de los datos de desplegamiento térmico en equilibrio (incluyendo los cinéticos) mediante un modelo mecánico-estadístico resultó en la estimación de una altura máxima de barrera (altura en el punto medio de desnaturalización) de aproximadamente 0.5 RT; es decir una barrera de energía libre minima. Los experimentos de fluorescencia en molécula únicas siguiendo el proceso mediante la señal de FRET y realizados a 288K y concentraciones de desnaturalizante cercanas al punto medio (C_m), confirmaron la existencia del comportamiento *downhill* con alturas de barrera estimadas de < 1RT en esas condiciones. Estos resultados demuestran la capacidad de un abordaje con múltiples sondas espectroscópicas para revelar las complejidades intrísecas al plegamiento de proteínas ultrarrápido y como una elección apropiada del modelo mecánico-estadístico para el análisis de los datos experimentales es capaz de conseguir una interpretación cuantitativa general del proceso de plegamiento de este tipo de proteínas.

Buttressing a New Paradigm in Protein Folding: Experimental Tools to Distinguish between Downhill and Multi-State Folding Mechanisms

By

Nagalakshmi, Tiruvarur Sooriyanarayanan

Dissertation submitted to the Faculty of Sciences of the
Autonomous University of Madrid, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Supervisors:
Professor Victor Muñoz
Dr. Mourad Sadqi

Dedication

To my parents

Acknowledgements

I thank my supervisor, Prof. Victor Muñoz, for his valuable guidance throughout my thesis work. He would not dare to initiate or do a research project from a different perspective, like my current research work, with a lot of enthusiasm, and it would probably take you the time until you get your first results in order to understand his perspective. He would also strongly encourage you to go ahead with that. I have also learnt a lot from the way he performs experimental data analysis and he would actually spend hours with you to actually do that and I have thoroughly enjoyed that. I have always felt comfortable in even asking stupidest questions to discussing sophisticated research topics with him. I would attribute my growth in the research field from being an undergraduate student to what I am now to him.

I must thank Dr. Mourad Sadqi, a senior scientist in Muñoz group who is also my co-supervisor, for he played a major role in getting myself settled in the laboratory when I joined and also for his guidance in biochemistry and in using bulk experimental techniques that are used in the research work. I must also mention that he is currently doing the project of atom by atom unfolding of engrailed homeodomain by NMR. This project would provide additional details and also complement the results that are shown in this thesis.

I would like to acknowledge, Dr. Michele Cerminara, another senior scientist in Muñoz group, for he literally acted as my another mentor in the group, especially when it came to performing kinetic and single molecule experiments. Instead of doing an experiment just one fine day with the mindset to expect a favorable result on the same

day, I have learnt to enjoy the process of performing experiments, whatsoever be the result obtained on a particular day, from him. I have also learnt a lot from the way he would help improvise the results in a step-by-step manner.

I acknowledge the laboratory of Dr. Jose Manuel Sanchez Ruiz from the University of Granada for performing the calorimetry experiments. I also acknowledge Dr. Athi Narayanan from IIT-Madras for helping me with Bayesian analysis.

I acknowledge the Spanish Ministry for a FPU predoctoral fellowship that provided me with a stipend during part of my stay in Spain, and Dr. Muñoz for providing me with a Marie-Curie PhD position through his grant that covered the expenses during the remaining part of my stay.

I thank Ignacio for translating the abstract and conclusions of my thesis to Spanish.

I take this opportunity to acknowledge in general other members of the Dr. Muñoz group.

I must also thanks to Dr. Raul Guantes, who is the coordinator of the biophysics doctorate program and also my tutor from UAM, for his constant help with procedures pertaining to the university.

I must thank all my wonderful flat mates over time, Sweety, Tania, Patricia, Luisa, Elena, Christina and Tiziana, who are also my friends, and I have shared many memorable moments with them. I must also thank the Indian junta in Madrid, for the good times I have had with them more especially during the festivity times, and in particular, Siya, who has become a great friend of mine.

Despite the fact that I am one of very few persons that end up doing PhD in Europe from my undergraduate batch while others ended up in U.S., and more so, being the only person from that batch in Spain, I have never missed most of my friends from my undergrad times for we have always been in touch and in particular I like to thank lokesh and sriram for they have always been there for me. I must also mention the support and love I have received from my childhood friends, Meenu and Ram, that have helped me to be sane over these times.

Finally, I want to thank my parents Sooriyanarayanan and Padmavathy, and my sister Iswarya, for their love, support and encouragement and more so for the faith they have always had in me.

Thank you every body and thank you for everything!

Table of Contents

Chapter 1: Introduction

1.1 Protein Folding Mechanisms

Protein molecules, long polymers of amino acids, are always in thermodynamic equilibrium, i.e. they constantly change their state by toggling between the native, biologically active, 3D structure and an ensemble of unstructured conformations. This process, known as protein folding, has important implications for the way proteins work in the cell and interact with one another. The (mis)folding of proteins has been implicated in several diseases such as Alzheimer's, Huntington's and Parkinson's. Protein folding is one of the few fundamental research problems that remain unsolved and characterizing the biophysical properties of proteins is an effective way to understand this process at a quantitative level. Understanding the rules that govern protein folding also has critical practical implications, as it would help predict the three-dimensional structural structure from the amino acid sequence and thus can directly read the mechanistic information from the corresponding DNA sequences, as well as designing new proteins.

In spite of numerous efforts during the last 5 decades, progress in the field has somewhat stalled, mostly because of an operational disconnect between the experimental and theoretical efforts. Due to the inherent complexity of the protein folding problem, theoretical and/or computational models must be parameterized empirically. At the same time, even the most sophisticated folding experiments do not provide mechanistic information directly, and thus need to be analyzed and interpreted using theoretical models. Folding experiments have been conventionally described in analogy to elementary chemical reactions: protein molecules must convert from the unfolded to the

native state by crossing one (two-state) or several (multi-state) large free energy barriers. This idea contrasts with the predictions from the energy landscape theory, which propose a new physical paradigm in which folding free-energy barriers are small, or even inexistent (downhill folding)[1], thus resulting in broad complex folding processes that resemble second-order phase transitions. Such downhill folding proteins can be of much interest because proteins that cross a single high free energy barrier will result in the mere observation of only the folded and unfolded ensembles, whereas the proteins that fold globally downhill or with a very small barrier(marginal barrier) will lead to observing the intermediate conformational ensembles by spectroscopic measurements, and will result in the movement of these ensembles in a continuous or semi-continuous way as the protein gets unfolded and would result in elucidating the complete folding mechanism of these proteins. In general, experimentalists have been extremely reluctant to apply the new theoretical ideas because at a first glance the conventional paradigm seemed capable of explaining the available experimental data. Such state of affairs is now starting to change after the efforts of a few experimental groups have lead to the experimental demonstration, that some (fast-folding) proteins that were originally thought to fold in a two-state fashion, in fact do so following a barrierless (downhill) process[2].

Analysis of the DSC thermal melting curves for a series of single domain proteins using a statistical mechanical model like Variable Barrier Model resulted in the direct estimation of barrier heights to folding for these proteins[3]. Based on barrier heights estimated, such single domain proteins were classified as globally downhill (zero barrier), marginal barrier also referred as downhill (< 2 RT barrier near characteristic/mid-denturing conditions) and two-state like proteins (> 4 RT near T_m). Proteins with barrier

heights between 2 RT and 4 RT are referred to as twilight zone proteins as their barrier heights are not significant enough to be called two-state proteins and these proteins could have the tendency to become either two-state or downhill proteins on very little perturbations.

BBL and gpW are two particular examples of downhill folding behavior. BBL was found to fold globally downhill, whereas gpW was found to fold over a marginal barrier. If a protein has a high folding barrier, there should be an overlap of extremely sharp unfolding behaviors studied by different spectroscopic techniques. In contrast to that, in cases of BBL and gpW[4], a broadness in the unfolding transition and a non-coincidence in the thermal unfolding between the unfolding curves obtained from different spectroscopic techniques that monitored different structural properties (global unfolding – DSC; tertiary structure – Fluorescence, Near UV CD; secondary structure – Near UV CD, FTIR) of the protein were observed and reported[2]. This was further illustrated by atom-by-atom thermal unfolding studies on BBL using a high resolution technique, NMR[5]. In this study, each amino acid residue acted as a different probe and there was a large spread of melting temperature between the unfolding curves from different amino acid residues. The average/global unfolding curve from this measurement overlapped with the unfolding curve obtained from CD measurement for this protein. Experimental signatures distinctive of a downhill folding mechanism based on the results obtained from the thermodynamic experiments have been reported[6,7].

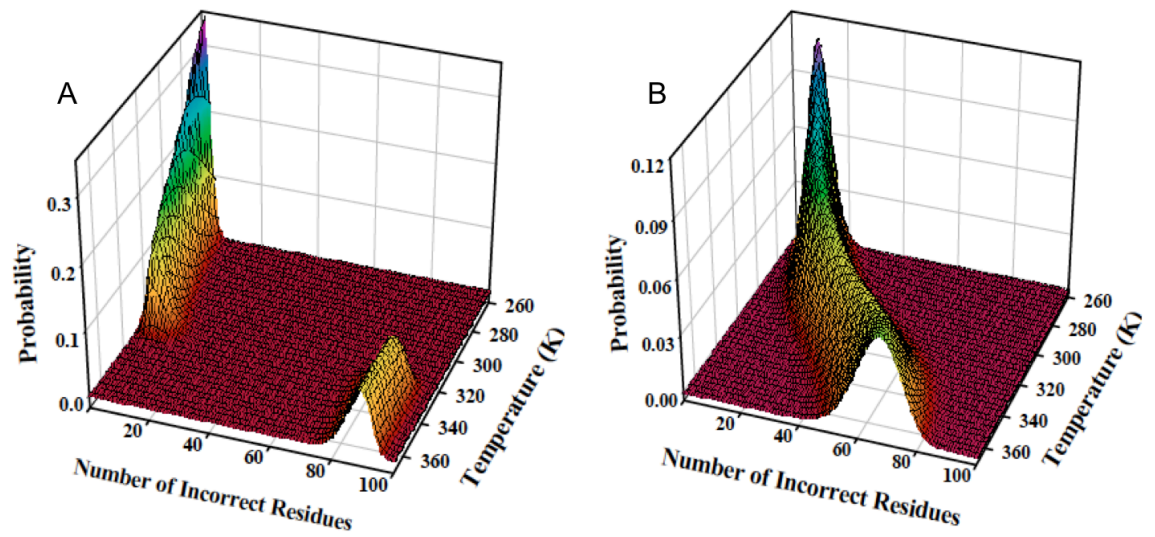


Figure 1-1: *Probability distribution of two-state (A) and downhill (B) folding scenarios at different temperatures as a function of the reaction coordinate, number of incorrect residues. Two-state folding shows clear separation of folded and unfolded conformations and changes in the folded and unfolded sub-populations with temperature, whereas global downhill folding results in continuous unfolding with temperature.Figure taken from the reference*[8]

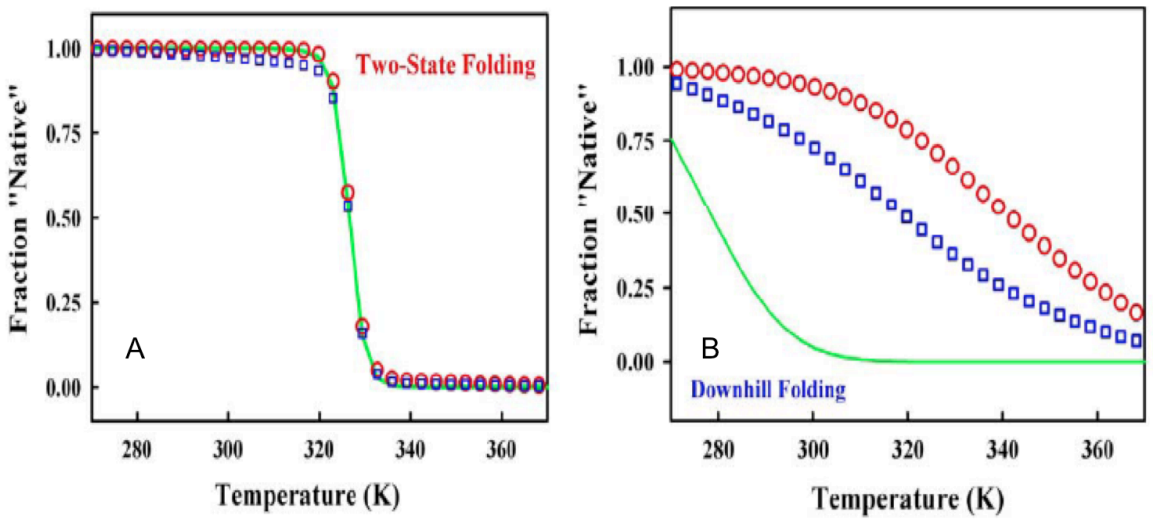


Figure 1-2: *Fraction native as a function of temperature from three different spectroscopic probes. Blue squares, red circles and green line represent the results obtained from three probes for two-state (A) and downhill (B) folding scenarios. Two-state folding shows sharp overlapping curves while downhill folding shows broad and non-overlapping curves. Figure taken from the reference*[8]

1.2 Fast Folding Proteins

Ultrafast T-jump kinetics help in measuring the protein folding rates in the order of few microseconds timescale[9-12]. How fast a single domain protein can fold or what is the protein folding speed limit can be discussed by first having an assumption that such a protein cannot fold faster than its constituent secondary structural elements, α-helix, β-sheet and loops. Loop forming rates[13,14] are less than 0.1 μs and many order of magnitude faster than α-helix and β-sheet of similar length. In the case of α-helix, 0.5 μs has been reported as the limit[15,16,17] and for β-hairpin, the spread is large and the rate is slower in comparison to the other secondary structural elements[18,19]. The reported rates vary between 0.8 μs for a hairpin called peptide I to 20 μs for an engineered N-terminal peptide from ubiqitin that forms the hairpin. Though experimental and theoretical works on single domain proteins approximated/predicted the protein folding speed limit[20] to be N/100 μs, studies on secondary structural elements implicated that an α-helical protein should fold faster than for β or αβ protein of the same length.

Theoretical study predicted the folding rates from the length of the protein. This was given by $\log(\tau_f) = N^{0.5}$, where τ_f is the folding rate[21]. Experimental data also revealed such a correlation between the folding rate and the length[22]. Another study reported that correlation exisiting between the barrier heights estimated by analyzing DSC melting curves by Variable Barrier model to the folding rates at 298 K[3]. As a consequence of all these studies, proteins that are small (<50 residues or in that range) and fold in microseconds are expected to have marginal folding barriers or no barriers at all, with small α-helical proteins having the possibility of being a downhill folder in

comparison to single domain proteins with other topologies. Such fast folding and downhill behaviors have been reported in the cases of i) alpha helical proteins/domains like BBL[23] and mutated version of lambda repressor[24]; ii) engineered version of WW domain[25], a beta sheet domain and, iii) α+β topology protein like gpW[4]. Probe-dependence in the kinetic amplitudes have also been reported[23].

1.3 Single Molecule FRET Studies of Protein Folding

Single molecule studies enable us to look at the distribution of conformation of biomolecules under study in comparison to the bulk thermodynamic and temperature jump kinetic measurements. FRET acts as a spectroscopic ruler as it can be converted to distance and have been in large used to track the conformational changes happen in the distance range of 10 to 80 Å[26,27]. First single molecule FRET study on protein folding was reported in the case of GCN4-P1peptide in which the folding was measured as a function of denaturant[28]. The experiment was performed on immobilized molecules. smFRET measurement on freely diffusing molecules was later performed on CI2[29] on account of eliminating the effects caused by immobilization on the measurements before. After this, smFRET was used to study the folding of cold-shock protein[30] (Csp*Tm*). Though free diffusion measurements help avoid any possible effects caused by immobilizing the molecules, the limitation of this measurement is that the maximum time a molecule can be observed (in a confocal volume) is restricted to 1 ms[31]. Thus, immobilization measurements have been widely employed to study the folding process in which case molecules can be tracked until the fluorophores get bleached. Three common ways the protein molecules are immobilized to the surface are: i) encapsulating the

molecules in a surface-tethered vesicle[32], ii) expressing the protein with histag[33] in which case the protein is attached to the surface functionalized with Ni^{2+}/Cu^{2+}, iii) protein molecules are directly immobilized to polyethyleneglycol coated surface by a biotin-streptavidin-biotin linkage[34]. Effect of linkers and immobilization must be taken care of anyways. If the dynamics of the molecule is fast, free diffusion measurements can still be very useful.

smFRET was used to study the folding of SH3, a two-state protein, and BBL, a global downhill folder. In the case of SH3[35,36], a bimodal distribution was observed at almost all concentrations of denaturant, except for native and denatured conditions. Also, the folded and unfolded conformations of SH3 increased or decreased depending on the concentration of denaturant without any significant movement of folded/unfolded peaks. In the case of BBL[35,37], a unimodal distribution was observed at all concentrations of denaturant (UREA), with the movement of distribution from higher FRET values to lower FRET values on increasing the concentration of UREA. These experiments were carried out at low temperature and using a specially developed photo protection cocktail[34-36] (Trolox and Cysteamine) to obtain high photon counts without bleaching and blinking. This allowed the use of 50 μs binning time that was necessary to analyze the photon trajectories of BBL and resolve the conformations as the reported relaxation rate of BBL at this experimental condition was ~ 150 μs.

smFRET studies also helped in the determination of protein folding transition path times of WW domain and protein GB1[38]. Transition path times can be very fast and experiments need to be performed under conditions in which the transition path can be observed/resolved by smFRET measurements. In this case, experiments were performed

at high glycerol concentration on immobilized protein molecules. The average transition path time for a protein at this experimental condtion was obtained by analyzing the photon trajectories by a three state model using the maximum likelihood method, where the intermediate state was chosen to represent the small region of transition between the folded and unfolded FRET values. This value was then calculated at zero glycerol concentration based on the linear correlation between viscosity and transition rates.

1.4 Engrailed Homeodomain

1.4.1 Sequence, Structure and Function of Engrailed Homeodomain

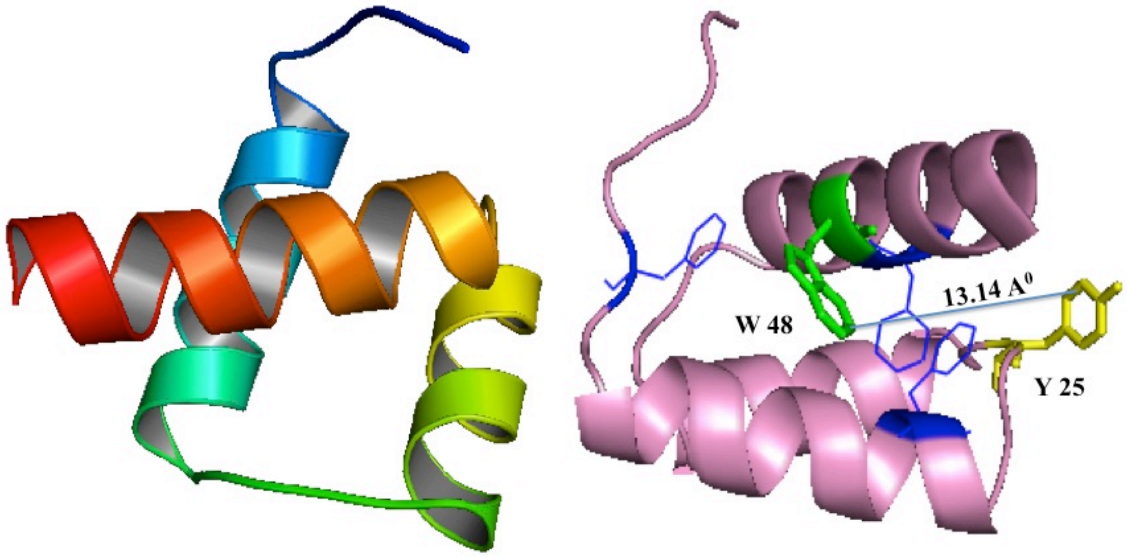


Figure 1-3: *Engrailed homeodomain structure (PDB ID: 1ENH). Figure on the right shows the distance between tyrosine and tryptophan present in the domain.*

Homeodomains are DNA-binding domains present in eukaryotes[39,40]. They consists of three helices. Engrailed homeodomain (EnHD) is one such a homeodomain from the fruit fly, *Drosophila Melanogaster*[41-43]. It is a small DNA-binding domain of 54 residues present as part of the large segmentation polarity protein in the fruit fly. This large protein, localized in the nucleus, has roles in transcriptional repression,

development of central nervous system and in specifying the body segmentation pattern[44,45].

SEQUENCE A - 54 Residues

RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI

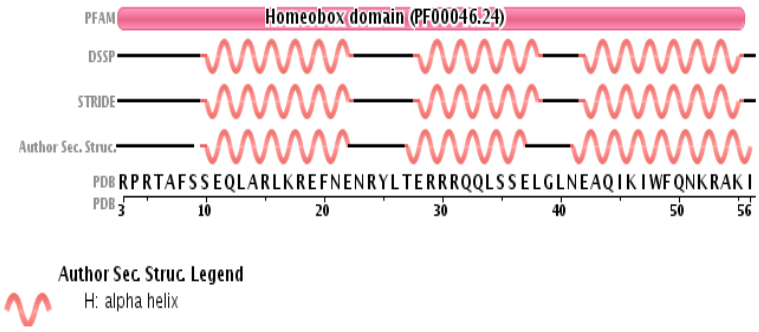


Figure 1-4: *Engrailed homeodomain sequence (top) (PDB ID: 1ENH). Figure on the bottom shows the secondary structural content predictions from different databases and from the author secondary structure*[41].

EnHD consists of five aromatic residues, three phenyl alanines, one tyrosine and one tryptophan.The core of EnHD structure has 4 aromatic residues F8, F20, W48, and F49 and such an aromatic core is conserved among many homeodomains. Especially, the position of tryptophan residue is retained in many known homeodomains and the residues 20 and 49 are always aromatics. EnHD bind the sequence (binding site) TAATTA with a high affinity[46,47]. N and C-termini residues become ordered upon binding to this sequence. There were no significant conformational changes observed in the helix 3,which is the recognition helix that binds to the major groove of the DNA, between the DNA-bound and the free structure of the protein. Based on the studies that were discussed in this chapter before, it is extremely likely that a small alpha helical domain like engrailed homeodomain can be a possible downhill candidate.

SEQUENCE B – 63 Residues

AC**EK**RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI**KKST**C

SEQUENCE B with Fluorescent-Labels

AC**EK**RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI**KKST**C

Alexa488 Alexa594

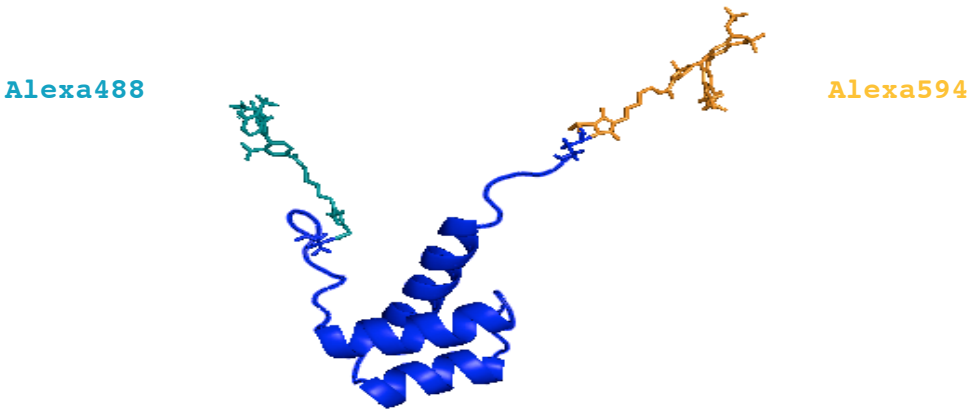


Figure 1-5: *Engrailed homeodomain sequence with added cysteines at the end (top) (PDB ID: 3HDD – for the amino acid sequence at the termini) followed by the sequence in which labels are marked attaching to cysteines. Figure on the bottom shows the structure of engrailed homeodomain covalently attached to Alexa 488 and Alexa 594 labels.*

1.4.2 Folding Mechanism of Engrailed Homeodomain – Research Problem

The proteins that fold in microseconds are expected to have marginal folding barriers(low barriers) or no barriers at all. Engrailed homeodomain, a 61-residue α-helical protein with a helix-turn-helix topology, folds in microseconds and the folding mechanism of this protein has been interpreted with a three-state model[49-57]. The conventional analysis of experimental data, as it has been done in the case of engrailed, by inherent assumes a large free-energy barrier between folded and unfoled states or between folded, intermediate and unfolded states and never estimates the barrier height(s) between these states. A consequence of this assumption is that no states/conformations

can be observed between the extreme states. But, the fast folding kinetics of this protein imply a low/inexistant barrier. This implication is very important as the presence of barrier-less/low barrier would lead to the observation of many intermediate states by spectroscopic measurements. With the conventional analysis of experimental data being unable to distinguish between downhill and multi-state folding of fast folding protein like engrailed, how can we know whether the folding mechanism of engrailed is three-state or multi-state or downhill ?

We propose to use a **multidisciplinary approach** in which the folding process of engrailed homeodomain will be investigated with **multiple spectroscopic probes** to monitor different structural properties of the molecule combining standard thermodynamic experiments with laser T-jump kinetic measurements using the custom-built in Nanosecond Resolution Temperature Jump instruments. In the case of performing nanosecond T-Jump experiments, the critical issue is to measure the relaxation rates and the kinetic amplitudes of all possible phases observed with all the probes. In these experiments, a downhill scenario will show quasi-equivalence of rates for all structural probes, but differences in the kinetic amplitudes, which will exhibit shifted maxima and partially overlapping curves. A multi-state process will show distinct kinetic phases with different rates and identical amplitudes. These thermodynamic-kinetic multi-probe experiments will be combined with differential scanning calorimetry. Finally, single molecule FRET experiments with microsecond resolution will be performed to resolve the distribution of unfolding behaviors in engrailed, and thus confirm the conclusions extracted from the previous analysis of bulk experiments.

1.5 Research Objectives and Chapter Summary

Research Objectives:

The specific objectives of the studies proposed here are:

1) To perform a complete thermodynamic and kinetic characterization of the protein by a multi-probe approach under different experimental conditions (temperature, chemical denaturants).
2) To look for signatures that can differentiate distinct folding behaviors by a detailed analysis of the transitions displayed by different experiments at different conditions (comparison of rates and amplitudes).
3) To assess the cooperativity exhibited by the folding process primarily by different scanning scanning calorimetry and other experiments.
4) Perform a series of microsecond-resolution single molecule FRET experiments on a variant of engrailed homeodomain labeled with a donor-acceptor pair of fluorophores.

All these experiments will lead to the determination of whether engrailed homeodomain protein folds downhill or multi-state, and more importantly, to develop a set of experimental and analytical methods that will be applicable to any other fast-folding protein. Such method will be instrumental for reinforcing the new physical paradigm in protein folding and will open a new avenue of experimental research in the field.

Chapter Summary:

Chapter 2 gives a brief overview of all the experimental methods used in this research. It also discusses various models that are used in the analysis of the experimental

data. Important thermodynamic and parameters that are extracted from the model are clearly specified. It also talks about the estimation of barrier height from different experimental results. Necessary sample preparation protocols and protein purification protocols are also given in this chapter.

Chapter 3 presents experimental results from equilibrium thermal unfolding measurements and analysis performed on those measurements. Equilbrium unfolding measurements from different spectroscopic techniques reveal heterogeneity in the unfolding behaviour of engrailed homeodomain based on two-state/first-derivative analysis. Fluorescence thermal unfolding data reveals complex unfolding mechanism. All the complex equilbrium thermal unfolding data are globally fit to a simple one dimensional free energy surface model – Mean Field Model. Extracting the barrier height from the global fit of the data using MF Model gives a barrier height of < 1 RT, clearly revealing a downhill folding mechanism for this protein.

Chapter 4 presents experimental results from double perturbation measurements from different spectroscopic probes. Two kinds of double perturbation measurements were performed i) thermal unfolding measurements over the entire range of denaturant by Far UV CD and, ii) chemical unfolding measurements over a series of temperature by Far UV CD and Fluorescence on the unlabeled protein, and by Fluorescence on the fluorescent-labeled protein. Results from each double perturbation measurement are globally fit by a two-state model. Results from the analysis clearly show deviation from two-state behavior and conform to the signatures of downhill folding mechanism.

Chapter 5 presents results from Infrared and Fluorescence-Temperature Jump kinetic experiments (thermal unfolding) and analysis performed on the same. Infrared

kinetic decays measured at two frequencies reveal single exponential kinetics whereas the spectral decays from Fluorescence-Tjump measurements reveal three significant components from SVD analysis and non-exponential kinetics. Results from the kinetics are globally fit to Mean Field Model by extending the analysis from the equilibrium and they conform to downhill mechanism but still accounting for the complexities observed in the kinetics.

Chapter 6 finally presents results from single molecule FRET measurements performed near chemical denaturation midpoint (C_m). Slow relaxation rate of engrailed at 288 K near C_m is ~ 127 μs. smFRET measurements near C_m at this temperature explored the conformational distribution in the folding transition path region of engrailed and revealed a bimodel distribution. Barrier height estimated by fitting the experimental data to 1DFES model used in conjunction with maximum likelihood method is < 1 RT. Results show the downhill nature of this protein.

Chapter 2: Materials and Methods

2.1 Expression and purification of unlabeled Engrailed Homeodomain without Cysteine at both ends (Sequence A)

SUMO-Engrailed homeodomain (EnHD) with the His-tag attached to the N-terminal of the SUMO fusion part, was obtained in a pSUMO vector (Top Gene Tech). The SUMO-EnHD fusion protein was expressed in *E.Coli* BL21 (DE3) Gold Strain. Cells were grown up to an O.D of 1.3 in 4L LB medium at 37^0 C and were induced with 1mM IPTG for 3 hours at 30^0 C. Cells were pelleted out by centrifuging for 30 min at 9000 rpm at 4^0 C and then re-suspended in lysis buffer (20mM Sodium Phosphate Buffer, 150 mM NaCl, 20mM Imidazole, 0.1 % Triton, 1mM PMSF at pH 7). Cells were lysed by passing through French Press thrice at a pressure of 1300 psi. The lysed cells were centrifuged for 1 hr at 30,000 rpm at 4^0 C. The fusion protein present in the supernatant was saturated with ammonium sulfate to 85%. The precipitated protein was resuspended in the binding buffer (20mM Sodium Phosphate Buffer, 150 mM NaCl, 20mM Imidazole at pH 7) and passed through 5mL His-Tag Crude Column (GE HealthCare). Pure fusion protein was obtained by eluting it using 20 mM Sodium Phosphate buffer containing 0.5 M Imidazole and 150mM NaCl at pH 7. After extensive dialysis against the binding buffer, the SUMO fusion part was cleaved by incubating the fusion protein with ULP1 protease (ULP1:Protein 1:100mg) for 2 hours at 37^0 C. Pure EnHD was obtained by passing the above reaction mixture through a C4 Reverse Phase Column (Higgins Analytical, Inc). For the Infrared measurements, in order to avoid signal contribution from TFA in the Amide I region, the second step of the purification was replaced by passing the cleavage reaction mixture through 5mL HIS-Tag Column (GE Health Care). Molecular mass of

the protein from both the purification was confirmed by Mass Spectroscopy (Mr = 6605.8 Da), and the protein obtained was > 99% pure. The protein samples were then lyophilized and stored in -20^0 C.

All the unlabeled equilibrium and kinetic measurements were performed using the protein sample obtained from this purification.

2.2 Equilibrium Thermal and Chemical Unfolding Measurements

2.2.1 Differential Scanning Calorimetry

Differential Scanning Calorimetric experiments (DSC) have been used to extract energetic information or thermodynamics of protein folding or protein interactions for almost four decades. In a typical DSC experiment performed to study protein folding, heat or energy is simultaneously introduced into the protein sample and the reference cell containing the buffer that is exactly used to prepare the protein sample. In other words, temperature in both the protein cell and the buffer cell is kept constant over time or the experiment is performed at the same scanning speed on both cells. While this is done, what actually is measured as energy of the protein sample is the amount of energy required to keep the temperature of the protein sample to be the same as that of the buffer sample. The amount of energy required depends on whether the process is exothermic or endothermic. An exothermic process will release heat and will result in a negative peak or transition in a DSC profile whereas an endothermic process will require more heat flowing into the cell containing the molecule of interest and will produce a positive peak in a thermogram. In a typical thermal unfolding measurement of a protein, the transition is characterized by a sharp endothermic peak at the thermal denaturation midpoint (T_m).

(In any case, a separate buffer/buffer scan is required in order to accurately estimate the heat capacity of the protein sample) The amount of energy required to increase the temperature (T) of one mole of protein sample by 1 K at constant pressure (P) is called heat capacity of the protein and this is measured vs. T. Enthalpy (ΔH) can be obtained from heat capacity by integrating the heat capacity over temperature.

$$C_P(T) = \left(\Delta H / \Delta T\right)_P$$
$$H(T) = \int_{T_0}^{T} C_P(T).dT + H(T_0) \qquad (2.1)$$

The partial molar heat capacity of a protein (experimental heat capacity, $C_{p,protein}^{\exp}$) as a function of temperature[58] (T) can be obtained by the following expression.

$$C_{p,protein}^{\exp}(T) = C_{p,buf}(T) \cdot \left(\overline{V}_{pr}(T) / \overline{V}_{buf}(T)\right) - \left(\Delta C_p^{app}(T) / m_{pr}(T)\right)$$
$$where \qquad (2.2)$$
$$m_{pr}(T) = \upsilon_{cell}(T) \cdot w_{pr}(T)$$

where $C_{p,buf}(T)$ is the heat capacity of the buffer, $V_{pr}(T)$ and $V_{buf}(T)$ are partial molar volumes of protein and buffer respectively, $m_{pr}(T)$ is the mass of the protein in the calorimetric cell, $\Delta C_p^{app}(T)$ is the difference between protein/buffer scan and the buffer/buffer scan, $\nu_{cell}(T)$ is the volume of cell and $w_{pr}(T)$ is the concentration of protein. $m_{pr}(T)$ is typically approximated to be constant. Partial volume of the protein can be calculated from the amino acid composition of the protein[59].

DSC Measurement and Sample Preparation. Calorimetry measurements were performed in a MicroCal VP-DSC differential scanning calorimeter (Northhamton, MA), fitted with a cell volume of ~ 0.5mL (Laboratory of José Manuel Sánchez Ruiz, Universidad de

Granada). Protein concentrations used were in the range of 0.9-2.75 mg/mL. DSC measurements of the protein were performed after extensive dialysis against the respective buffer (#) and the heat capacity of the protein was obtained after subtracting for the buffer. Measurements were done at two different scanning rates, 90K/hr and 200K/hr, and the reversibility of protein was checked after heating to ~ 398K for both the scanning speeds. Thermal unfolding measurements performed at a series of concentration of protein at the scanning speed of 1.5 K/min (90K/hr) were averaged and then taken for further analysis. Experimental error at every temperature was calculated from these independent measurements made at different protein concentrations.

(#)Experimental Condition: Buffer and Protein Concentration Determination

All the labeled and unlabeled protein samples for all the measurements were prepared in a 20mM Sodium Acetate buffer with 100mM NaCl, at pH 5.5. For chemical denaturation measurements performed using urea, samples were prepared in the same buffer with the respective concentration of urea. Concentration of urea was determined by refractive index (Abbe Refractometer). Protein concentrations were measured using a Cary 100 Bio UV-Visible Spectrophotometer. The extinction coefficient used for the unlabeled protein was 6970 $M^{-1}cm^{-1}$ [60].

2.2.2 Circular Dichroism

Circular Dichroism is an absorption spectroscopy that measures the difference in the absorption between left and right circularly polarized light of optically active chiral molecules. The differential absorption is given by the following expression.

$$\Delta A(\lambda) = A_L(\lambda) - A_R(\lambda) = [\varepsilon_L(\lambda) - \varepsilon_R(\lambda)] \cdot l \cdot c = \Delta\varepsilon \cdot l \cdot c \quad (2.3)$$

where A_L and A_R are absorptions corresponding to left and right-handed circularly polarized light and ε_L and ε_R are the corresponding ellipticity values and $\Delta\varepsilon$ is the differential molar circular dichroic extinction coefficient. $\Delta\varepsilon$ is related to molar ellipticity ([Θ]) as: $[\Theta] = 3300 \times \Delta\varepsilon$. In a typical CD experiment, [Θ] is measured in milli degrees. The recorded values in milli degrees is converted to mean residue ellipticity as $[\Theta]_{MRE}$ in deg.cm^2.dmol^{-1} for a protein sample as follows.

$$[\Theta]_{MRE} = \Theta / (10 \cdot l \cdot C \cdot N) \quad (2.4)$$

where [Θ] is the measured ellipticity in milli degrees, *l* is the path length of the sample in cm, C is the concentration of the protein sample in moles/liter (M) and N is the number of peptide bonds present in protein.

Secondary structural content of a protein is monitored in the Far UV region (190 – 250 nm). Peptide bond present in the proteins is the principle absorbing group in this region and yields characteristic spectra for α-helix, β-sheet and random coil as depicted in the figure (Figure 2-1). CD signal in the Far UV region can also have minor contributions from aromatic amino acid residues present in the protein. An α-helix spectrum typically has two negative minimums near 222 nm and 208 nm and one positive maximum near 190 nm. The minimum at 222 nm is due to $n \rightarrow \pi^*$ molecular orbital transition and the negative and positive values near 208 nm and 190 nm are due to parallel and perpendicular components of $\pi \rightarrow \pi^*$ molecular orbital transition coming from the delocalized electrons of the peptide bond. Similar transitions produce a negative band near 215 nm and a positive band near 198 nm for β-sheet. In the case of random

coil, these transitions produce characteristic negative band near 230 nm and a positive band near 195 nm.

A CD spectrum of a protein can be assumed to be an additive spectrum of these secondary structural contents and can be represented as

$$[\Theta]_{MRE} = f_\alpha \cdot [\Theta]_\alpha + f_\beta \cdot [\Theta]_\beta + f_R \cdot [\Theta]_R \tag{2.5}$$

where $[\Theta]_\alpha$, $[\Theta]_\beta$ and $[\Theta]_R$ are the mean residue ellipticity values for α-helix, β-sheet and random coil and f_α, f_β and f_R are the corresponding fractional secondary structural content.

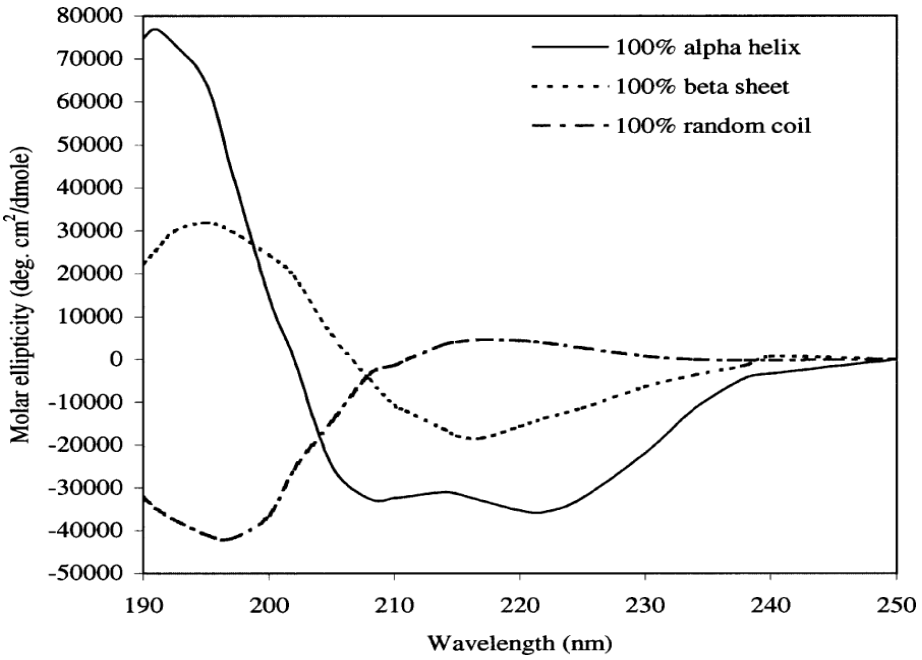


Figure 2-1*. CD spectra of different secondary structural elements. Figure adapted from the reference*[61]

For an all-helical protein, fractional helix content at room temperature can be calculated using the following expression

$$f_\alpha = \left([\Theta]_{208nm} - 4000 \Big/ 33000 - 4000 \right) \tag{2.6}$$

where $[\Theta]_{208nm}$ is the mean residue ellipticity at 208nm. Similarly, mean residue ellipticity value at 222 nm can also be a good estimate of fractional alpha helical content for an all-α protein.

In the Near UV region (250-300 nm), aromatic amino acids like tyrosine (Y), tryptophan (W) and phenylalanine (F) contribute to the CD signal. Due to large extinction coefficient of tryptophan and tyrosine, they come out as strong CD signal in this region when compared to phenylalanine. Tryptophan produces a peak near 290 nm with two more fine structure between 290 and 305 nm[106]. Tyrosine produces a peak between 275 nm and 282 nm and phenylalanine, a sharp fine structure between 255 and 270 nm. Disulphide bonds between thiol groups in Cysteines also contribute to CD signal in the 240-290 nm regions. Dihedral angle of the disulphide bond essentially comes out as CD signal. Basically, CD signal in the near UV region monitors tertiary environment of the proteins and the signal is typically very weak when compared to Far UV region. CD signal of a protein in the near UV region typically has convoluted effects from all the three amino acid side chains. As these amino acid side chains (primarily, aromatic) monitor the tertiary environment of the protein, the CD signal in this region is affected by charged residues present in the surroundings or the solvent, polarizability, hydrogen bonding and the extent these amino acid side-chains are exposed to the solvent.

CD Measurements and Sample Preparation. CD measurements were carried out in a Spectropolarimeter from Jasco (J-815) equipped with a Peltier cuvette holder. Protein concentration used for the Far UV CD measurement and for double perturbation experiments was 40 μM. For the chemical unfolding experiments performed at a series of

temperature between 5^0 C to 35^0 C using Far UV CD, protein samples were used at a concentration of 41 μM. Sample was prepared in a quartz cuvette of 1 mm path length. Measurement parameters used for these experiments were bandwidth of 2nm, response time of 16s, scanning rate of 10 nm/min and resolution of 1nm in a continuous scanning mode. Protein sample was equilibrated for 2 minutes at every temperature. Protein concentration used for the Near UV CD measurement was 114 μM. In the case of Near UV CD measurements, same parameters are used, but the sample was prepared in 1 cm path length quartz cuvette

2.2.3 Fourier Transform Infrared Spectroscopy

FT-Infrared spectroscopy is an absorption technique that measures the absorption of IR radiation by a sample and has long been used to measure secondary structural content of proteins. Basically, it is a vibrational absorption spectroscopy that measures the absorption from the ground vibrational state to a higher vibrational state when a ray of IR light passes through the sample. Every electronic state (in a molecular orbital) has several vibrational states and excitation of molecules from the first to nth vibrational levels produces several absorption bands. Typical molecular vibrations include stretching (symmetric and asymmetric), bending and twisting. In the case of proteins, vibrational motions of peptide bond produce several absorption bands and it has been reported to produce nine characteristic absorption bands (Amide I to VII, Amide A and B region). These vibrational motions of peptide bonds include C=O, CN and N-H stretching, C=O, N-H and OCN bending, in-plane N-H bending coupled to stretching motions, out-of-plane N-H and C=O bending and torsional motions. Bands in the Amide I region are

caused primarily by C=O stretching and also by C=O stretching coupled to in-plane N-H bending. This happens in the 1600-1700 cm^{-1} region. Bands in the Amide II and III regions are caused by CN stretching and N-H bending in the regions 1480-1575 cm^{-1} and 1230-1300 cm^{-1} respectively. NH stretching also produces vibrational bands centered around the frequencies 3300 cm^{-1} and 3100 cm^{-1}, which are called Amide A and B bands.

Bands in the Amide I region typically are directly related to secondary structural content of proteins. Peptide bonds are involved in hydrogen bonding within each secondary structural element. The orientation and the strength of hydrogen bonding between peptide bonds differ between secondary structural elements. Thus, the peptide bond vibrational motions coupled to these hydrogen bonds give rise to several absorption bands that are characteristic of different secondary structural elements within the Amide I region. Typically, water has a strong absorption band in the Amide I region and hence masks the FT-IR spectrum in this region. Thus, protein samples in buffers prepared with deuterated water are preferred. Strategies to deconvolute the spectrum prepared in water also exist but for convenience, a deuterated water solvent has been used. (Absorption bands assigned for different secondary structural elements in both H_2O and D_2O have been reported). A typical FT-IR spectrum is a convoluted spectrum of different secondary structural elements and hence has to be decomposed to calculate the percentage of secondary structure present. Common deconvolution procedures include: i) a spectrum can be split into 'n' Gaussian or Lorentzian peaks directly, ii) second derivative of the spectrum by knowledge-based approach, and iii) Fourier-Self Deconvolution (FSD). All these de-convolution procedures typically come as part of the software that acquires FT-IR spectrograph. These procedures enable us to obtain more quantitative information

about protein secondary structures. The following Amide I absorption bands for proteins are for the samples prepared in D_2O. A typical α-helical spectrum is centered around 1653 cm^{-1}. α-helix also produces a band 20 cm^{-1} below the main peak, depending upon whether the alpha helix is buried or exposed. β-sheet produces many absorption bands in this region and they are around 1624 cm^{-1}, 1631 cm^{-1}, 1637 cm^{-1} and 1675 cm^{-1}. β-turns have bands around 1663 cm^{-1}, 1671 cm^{-1}, 1683 cm^{-1}, 1689 cm^{-1} and 1694 cm^{-1}. A random coil's feature is around 1646 cm^{-1}. These are absorption bands characteristic of different secondary structural elements[62,63] and depending on proteins, these bands might vary more or less significantly. Some amino acid side chains also produce absorption bands in the Amide I region[64].

In a typical protein folding experiments, a buffer spectrum is acquired at every temperature as buffer components in the Infrared region have strong temperature dependence. Thus, an FT-Infrared spectrum of a protein at every temperature is subtracted from the buffer spectrum acquired at the same experimental condition. The resultant FT-Infrared spectrum of a protein at a particular condition is calculated as

$$Abs_{protein} = -\log_{10}\left(T_{protein}/T_{buffer}\right) \tag{2.7}$$

where $T_{protein}$ and T_{buffer} are protein and buffer transmissions.

FT-IR Measurement and Sample Preparation

The equilibrium infrared unfolding measurements were performed in a Jasco FTS-300R IR Spectrometer. Sample solution was mounted in a cell formed by two disks of CaF_2 separated by a 50μm teflon spacer, held in a cell holder from Harrick scientific that enables to thermalize the samples in the range between 5 ^{0}C and 95^0 C. For equilibrium

infrared measurement, protein was deuterated by cycles (3 cycles) of thermal treatment in D_2O followed by lyofilization; samples were then prepared in the same buffer described above, but using 99.9% deuterated water (Sigma Aldrich) at pD 5.5. $pH_{reading}$ value of the sample/buffer prepared in a deuterated water was corrected for the solvent isotope effect on the glass electrode reading or for the D^+ ion concentration[107] as: $pD = pH_{reading} + 0.40$. The concentration of protein was 0.9 mM for equilibrium FT-IR measurements. Measurement parameters used for spectral acquisition in the range 700 cm^{-1} - 4000 cm^{-1} were: 100 accumulations with a spectral resolution of 1 cm^{-1}. Samples were equilibrated for 4 minutes at every temperature.

2.2.4 Steady-State Fluorescence

Fluorescence is an emission property of the molecule. When a ray of light is passed through a sample, it absorbs the photon/light that matches with the energy difference between ground electronic singlet state and excited electronic singlet states. This happens within 10^{-15} s. Every electronic state has many vibrational states. When molecules absorb light, they are usually excited to higher vibrational levels of the first or second singlet energy state. In the excited electronic singlet states, molecules relax back to their ground vibrational levels and further relax to the ground vibrational level of the first singlet-excited state. The later process is called internal conversion and occurs on the order of 10^{-12} s. When a molecule is in this state, it emits a photon and comes back to the ground state. This emission is called 'Fluorescence' and this occurs in a time scale of 10^{-9} - 10^{-7} s. A molecule at excited singlet state can also go to a triplet state in a non-radioactive way. This process is called intersystem crossing. When the molecule goes to

the triplet state, it again relaxes back to its ground vibrational state and emits a photon and comes back to the ground state. This process is called 'Phosphorescence'. A triplet to a singlet state or vice versa is forbidden and thus this process is longer than fluorescence and occurs on the order of several of seconds. All these are represented in the 'Jablonski Diagram' (Figure 2).

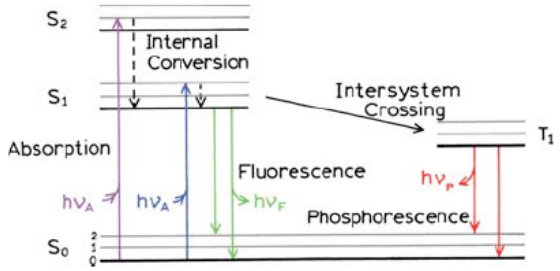


Figure 2-2. *Jablonski Diagram. Figure adapted from the book 'Principles of Fluorescence Spectroscopy'*[102].

When a molecule emits a photon, the wavelength of the emitted photon is longer than that of the incident photon. As the molecule rapidly relaxes to the lowest vibrational state of the first excited state, fluorescence emission is independent of excited wavelength. Quenching, energy transfer and solvent interactions affect fluorescence of a molecule. Fluorescence is extremely sensitive to surroundings and thus,can be used to monitor the tertiary structural content of the protein in the unfolding process.

Quantum Yield of a fluorophore is a ratio of number of photons emitted over the number of photons absorbed. Tryptophan, tyrosine and phenylalanine are naturally present fluorophores in proteins. These amino acid residues can be excited at 280 nm, 274 nm and 257 nm respectively and the resultant fluorescence spectrum will have emissions centered around 343 nm, 303 nm and 282 nm respectively. As tryptophan has

the highest extinction coefficient, it has a higher quantum yield of 0.2 in comparison to 0.14 and 0.04 of tyrosine and phenylalanine respectively near neutral pH. Fluorescence lifetime of a fluorophore is the amount of the time a fluorophore spends in its relaxed first excited singlet state. If a fluorophore has a single exponential relaxation, it is the time it takes to come to 1/e of the value when it relaxes from the first excited state to its ground state.

The lifetime (τ) of a fluorophore is given by

$$\tau = \left(1 \Big/ k_{fl} + \sum k_{nr} \right) \tag{2.8}$$

where k_{fl} is the emission rate of the fluorophore and the second term in the denominator represents the sum of the rates of non-radioactive decays, each represented by k_{nr}.

Quantum yield (Q.Y.) is related to k_{fl} and the sum of k_{nr} by the following expression.

$$\text{Q.Y.} = \left(k_{fl} \Big/ k_{fl} + \sum k_{nr} \right) \tag{2.9}$$

The lifetime of a fluorophore in the absence of any other non-radioactive processes is called 'intrinsic life time of a fluorophore' and can be calculated by the measured life-time (τ) and the quantum yield. But processes such as phosphorescence and quenching of fluorescence by nearby residues affect this calculation.

Quantum Yield of a protein ($\text{Q.Y.}_{\text{protein}}$) is calculated by the following expression.

$$\text{Q.Y.} = \text{Q.Y.}_{NATA} \cdot \left(A_{protein} / A_{NATA} \right) \cdot \left(Abs_{NATA} / Abs_{protein} \right) \tag{2.10}$$

where $\text{Q.Y.}_{\text{NATA}}$ is the quantum yield of NATA (N-Acetyl Tryptophamide) at pH 7 which is 0.13 at 298 K, $A_{protein}$ and A_{NATA} are the area under the fluorescence emission spectrum

of protein and NATA respectively, and Abs_{NATA} and $Abs_{protein}$ are the corresponding absorbance values at the same excited wavelength. Protein and NATA samples have to be measured at the same experimental condition. Quantum yields of fluorophores have strong temperature dependence and tryptophan quantum yield is heavily affected by the pH of the solution.

Steady-State Fluorescence and Sample Preparation

Fluorescence measurements were done in a Fluorolog-3 spectrofluorimeter (Jobin Yvon) equipped with a thermostatted sample holder. For thermal unfolding measurements, protein samples were excited at 280nm and the emission spectra were collected from 290 nm to 550 nm, every 5^0C from 0^0C to 95^0C. For this measurement, 18μM protein sample was used. For chemical unfolding double perturbation experiments, unlabeled protein samples were excited at 280nm and the emission spectra were collected from 290 nm to 550 nm, every 10^0C from 5^0C to 35^0C. For these measurements, unlabeled protein samples were prepared at a concentration of ~7.3 μM. Samples were prepared in a quartz cuvette of 1 cm path length. Measurements were acquired using the following parameters: slit width of 5 nm for excitation and emission slits and integration time of 0.25 s per nm. Samples were equilibrated for about 4 minutes at a particular temperature before acquiring measurements.

2.3 Temperature Jump Kinetic Measurements

Temperature Jump kinetic measurements are primarily used for studying ultra fast folding proteins that folds on the order of microseconds[108]. Laser Induced Temperature Jump kinetics is a perturbation technique in which a protein/buffer sample kept at a particular temperature is heated by a pump laser pulse by the excitation of the vibrational modes of solvent-water molecules that spontaneously heats up the sample locally. This occurs spontaneously as vibrational-relaxations occur in the order of picoseconds. The size of temperature jump produced depends on the energy of the laser heating pulse. It can also be affected by the amount of solvent molecules present in the path length.

What exactly happens to the sample in a typical Laser Induced T-Jump measurement? First, the sample's temperature spontaneously rises to the final temperature of the T-Jump because of solvent molecules. At this point, the protein is still in its initial probability distribution corresponding to the initial temperature of the T-jump and the protein still needs to respond to the increase in the temperature of the solvent. Depending upon the kinetics of the protein, the protein then starts to redistribute itself to its final probability distribution and reaches a signal value corresponding to the final temperature of T-Jump. Now, the solvent slowly cools down to its initial condition, that is, to the initial temperature of the T-Jump. A probe laser completely monitors the signal of the protein sample, be it absorbance or transmission or fluorescence emission.

For the current studies, both Laser Induced Infrared and Fluorescence Temperature-Jump kinetic measurements were performed. Both the apparatus were set up in the lab by our colleague Dr. Michele Cerminara

2.3.1 Infrared Temperature Jump Kinetics

Experimental Setup

A pump (Nd:YAG) Laser beam (Litron Nano L-10) emitting at 1.064 μm and shifted to 1.097μm with a Raman Cell filled with H_2 gas, was used to produce temperature jumps of ~ 11 K on the protein and the buffer samples by exciting the vibrational overtones of D_2O. H_2 gas shifts the incident wavelength to 1.097μm where the absorption of H_2O is very strong. H_2O also has strong absorption signals in the Amide I region where the signal needs to be probed. Thus, this set up is suitable for an Infrared T-Jump measurement using D_2O as solvent. The effect of the temperature jump was probed by measuring the transmittance at 1646 cm^{-1} and 1636 cm^{-1}, using a tunable quantum cascade laser (Light Age, Inc) as probe beam. The transmittance from the sample was detected by a nano second resolution IR detector. The sample solution was placed in between two MgF_2 windows separated by a 50μm Teflon spacer and placed in a custom designed sample holder in which the cell was thermostatted at the initial temperature before the T-jump. The signal is recorded by an Oscilloscope (Voltage vs. time) and is interfaced with the computer through a Lab View program.

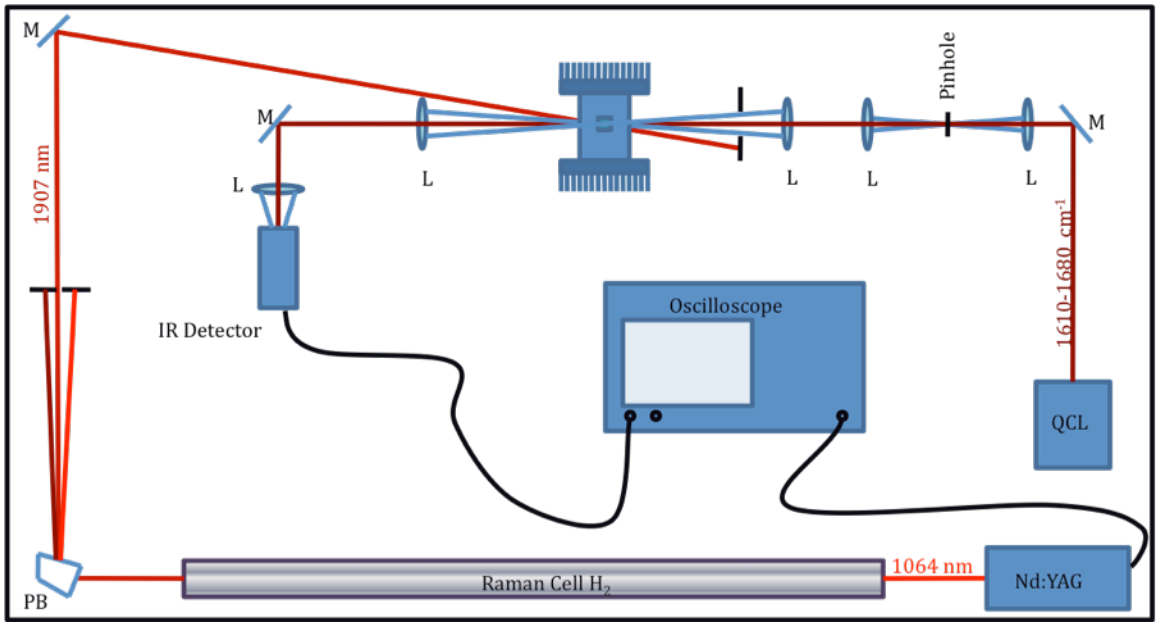


Figure 2-3. *Schematic of Infrared Temperature Jump Apparatus*

Temperature-Jump Calibration and Sample Measurements

In the case of Infrared, calibration of temperature jump was made using the buffer decays. Voltage transmission (at negative time scales) before the temperature jump was measured at every (initial) temperature. Thus, the voltage values for every temperature were known and thus a calibration graph was obtained. The voltage values just after the laser heating pulse were measured. These final voltage values were used to obtain the final temperature of the T-Jump using the calibration graph. In the measurements made, two frequencies were used to probe protein kinetics. Temperature jump calibration was made at two frequencies 1646 cm^{-1} and 1636 cm^{-1}. About 11.2 K and 10.8 K temperature jumps were produced at the frequencies, 1646 cm^{-1} and 1636 cm^{-1} respectively.

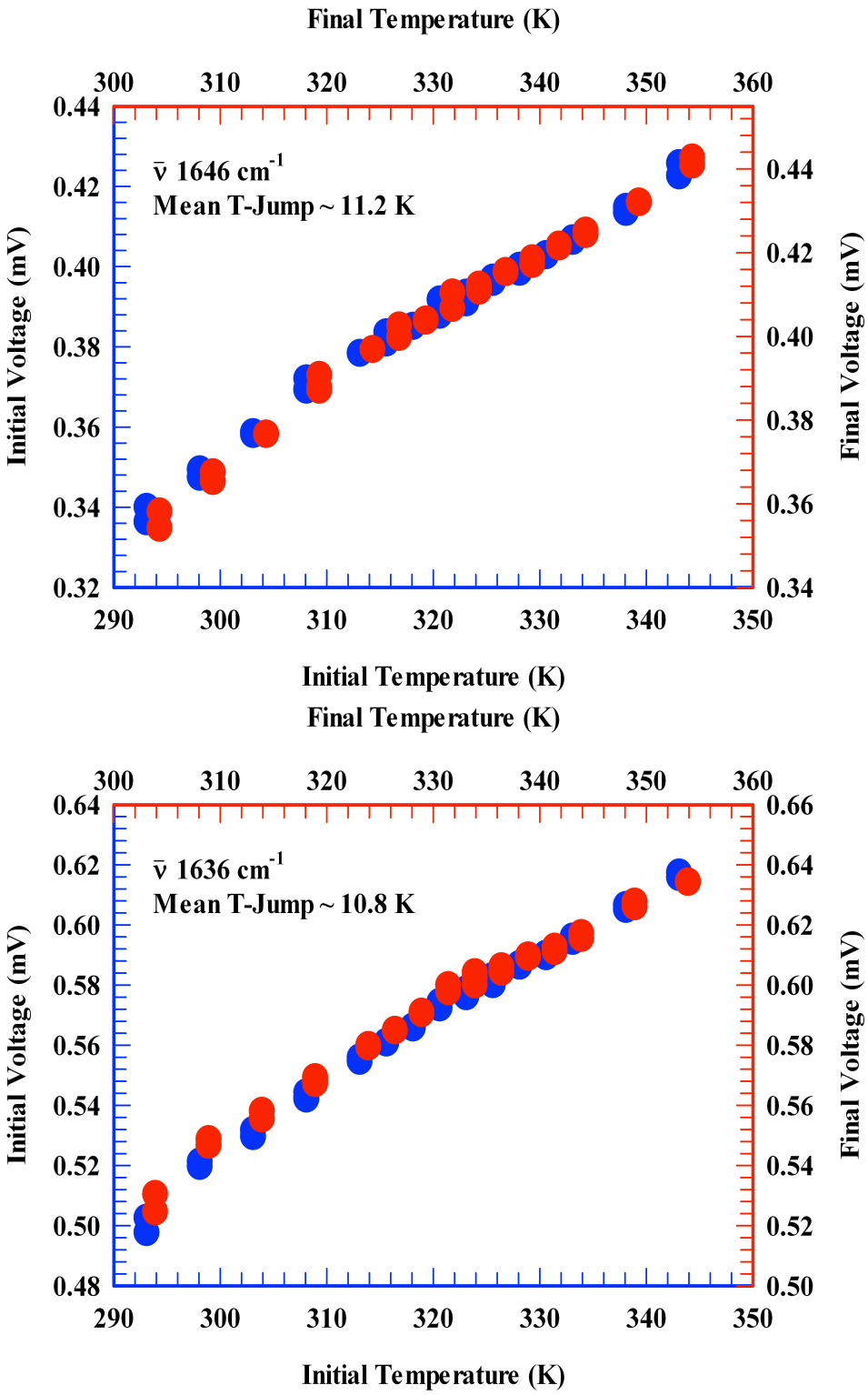


Figure 2-4. *Temperature Jump Calibration Curves. Upper and lower plots show the calibration curve for the frequencies 1646 cm^{-1} and 1636 cm^{-1} respectively. Voltages corresponding to the initial temperature of the T-Jump are shown as blue circles. Initial voltage vs. temperature serves as a calibration graph and temperatures corresponding to the voltage values after the temperature jump were extrapolated from the calibration graph.*

Protein sample decays and buffer decays were obtained at a series of temperatures. Protein relaxation (absorbance decay) was then obtained by,

$$Abs_{protein}(t) = -\log_{10}\left(T_{protein}(t)/T_{buffer}(t)\right) \quad (2.11)$$

where $T_{protein}(t)$ and $T_{buffer}(t)$ are protein transmission and buffer transmission decays. Typically, if the decays were single exponential, the cooling of the sample was fit by fitting the decays using a single exponential plus a drift line.

Sample Preparation

For infrared kinetic measurement, protein was deuterated by three cycles of thermal treatment in D_2O followed by lyofilization; samples were then prepared in 20 mM acetate buffer with 100 mM NaCl buffer as described above, but using 99.9% deuterated water (Sigma Aldrich) at pD 5.5. The concentration of protein used was 0.57 mM for the kinetic measurements.

2.3.2 Fluorescence Temperature Jump Kinetics

Experimental set up

A pump (Nd:YAG) Laser beam (Litron) emitting at 1.064 μm and shifted to 1.561 μm using a Raman Cell filled with D_2 gas, was used to produce temperature jumps of about 6.8 degrees. The shifted wavelength excites molecules of water and thus, heats up the sample. The fluorescence of the protein, used as a probe of the protein relaxation following the temperature jump, was excited using the 4th harmonic of a Nd:YAG laser (Continuum Minilite II) at 266nm , shifted to 288nm to reduce photo damage by using a Raman Cell filled with methane gas. The emission spectra were collected with a CCD camera (Princeton Instrument Pixis 100B) coupled to a spectrograph (Princeton

Instruments Acton Spectrapro-2150I). The delays between the pump and probe lasers were set using a digital delay generator (Stanford Research Systems DG535) and measured by a digital counter (Agilent technologies 53132A), with a jitter between the two pulses of about 1 ns.

Protein samples were put in a quartz cuvette with a path length of 0.5 mm and placed in a custom-designed sample holder in which the cell was thermostatted at the initial temperature before the T-jump.

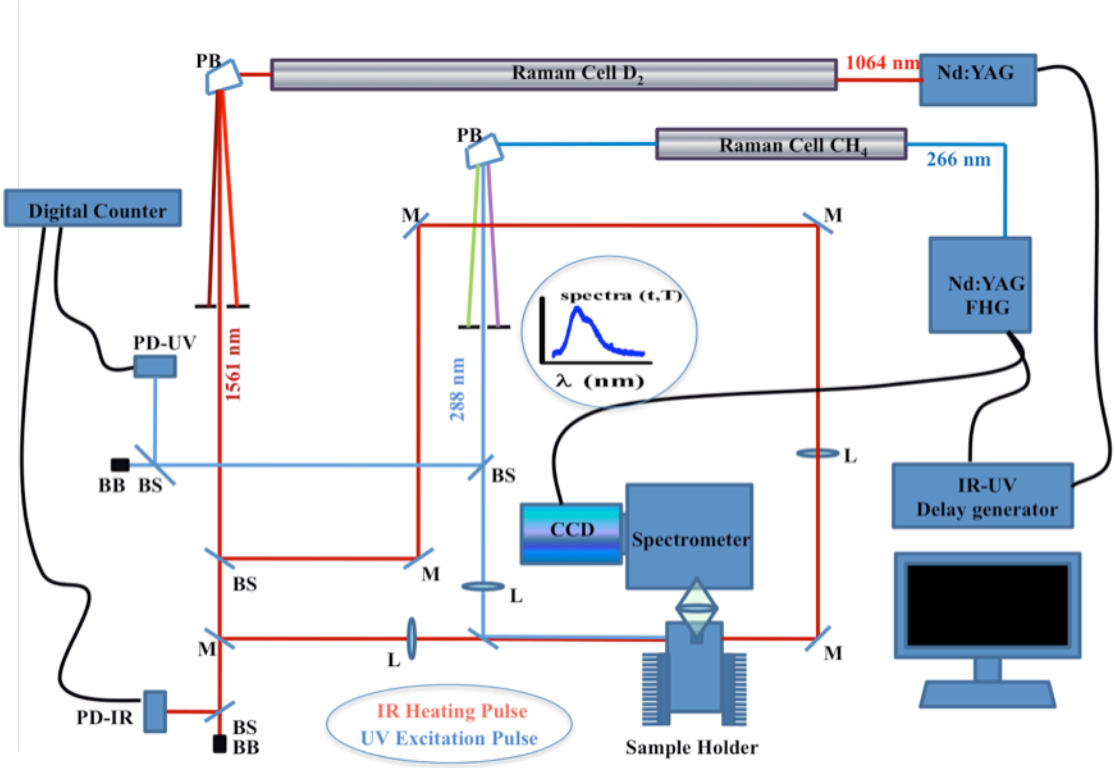


Figure 2-5. *Schematic of Fluorescence Temperature Jump Apparatus*

Temperature Jump Calibration and Sample Measurements

In the case of fluorescence temperature jump, calibration of temperature jump was made using NATA. NATA temperature ramp was made at equilibrium and quantum

yields were obtained at every temperature using the standard value of 0.13 at 298 K at pH 7. Quantum yield of NATA decreased with increase in temperature. Upon temperature jump, the quantum yield/fluorescence emission of NATA got decreased and again cooled back, after 1 ms, to the base temperature of T-jump. NATA cooling curves were obtained at a series of base temperatures (initial temperature of the T-jump). The fluorescence emission values before and after cooling were taken at every base temperature. Final temperatures of T-jump were obtained by extrapolating the emission values before cooling to the values after cooling. In our measurements, a temperature jump of ~ 6.8 K was produced on the sample. Figures 2-6 and 2-7 show the NATA calibration curve and a representative NATA cooling curve at 288 K respectively.

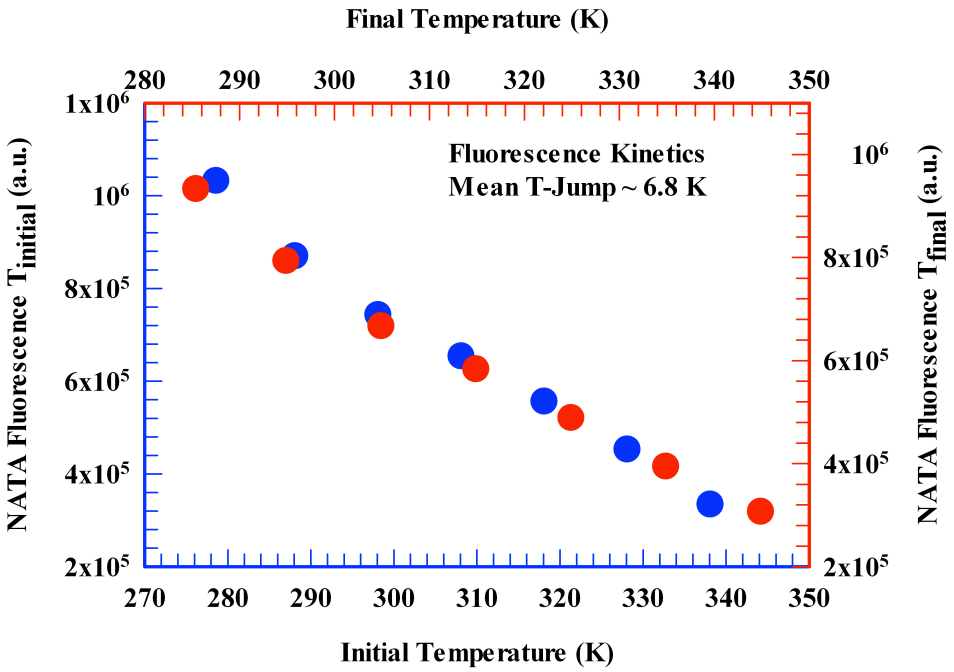


Figure 2-6. *NATA Calibration Curve*

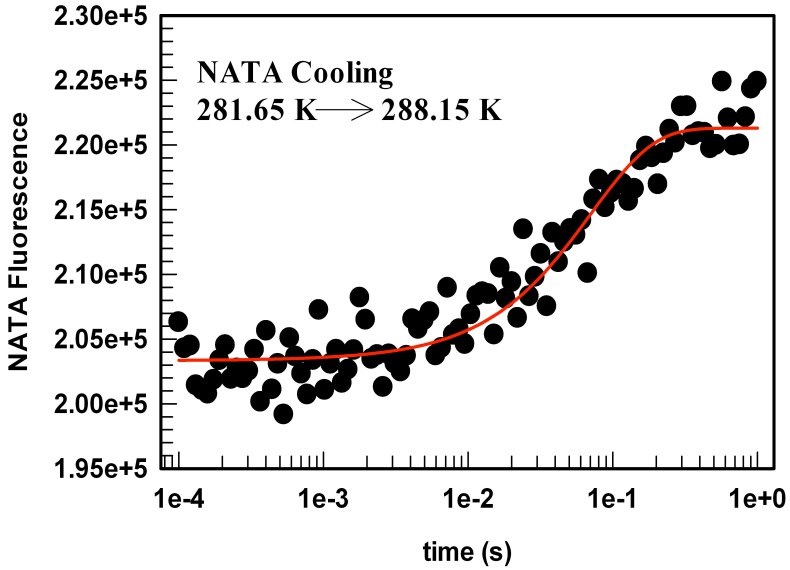


Figure 2-7. *Fluorescence T-Jump. Representative NATA Cooling Curve. Emission of NATA decreases with temperature jump and then increases upon cooling to initial temperature.*

Spectral decays were acquired for protein samples for a series of temperature. Area under the spectra or the fluorescence emission (a.u.) obtained at the base temperatures for the protein was normalized to the corresponding quantum yield values calculated from the steady-state measurements. The entire spectral decay was then represented in terms of the quantum yield (Figure 2-8). Once all the experimental spectral decays obtained at all the temperatures were rescaled based on the quantum yield, a global SVD analysis of the entire dataset was performed. The decays were then analyzed fitting to a single, double, stretched exponentials and/or using an appropriate statistical-mechanical model depending upon the experimental results obtained.

Sample Preparation

For Fluorescence T-Jump measurements, protein samples were prepared in the same buffer, at a concentration of ~137 μM. NATA samples used for the calibration of T-Jump

were in the concentrations of 200-250 μM. NATA samples were prepared in 20 mM Phosphate buffer at pH 7.

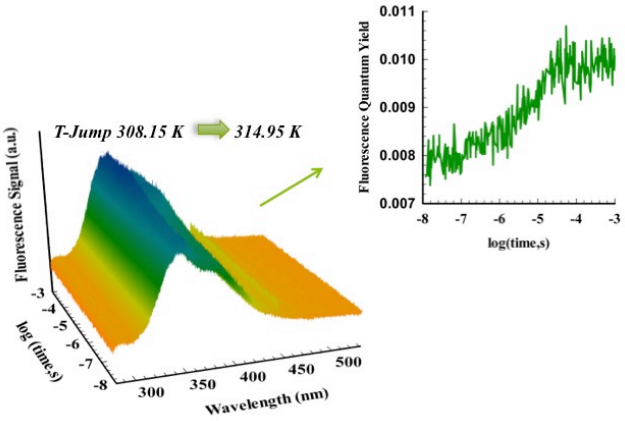


Figure 2-8. *A representative experimental spectral relaxation for the engrailed homeodomain at 315 K and the corresponding normalized decay represented in terms of the quantum yield of the protein.*

2.4 Preparation of fluorescent-labeled Engrailed homeodomain

2.4.1 Expression and purification of unlabeled Engrailed Homeodomain with Cysteine at both ends for fluorescent labeling

Engrailed homeodomain with cysteines on both ends (sequence B) was expressed in E.Coli (BL21 (DE3)). The cells were grown up to an O.D of 1.2 in a 4L LB medium at 37^0 C and were induced with 1mM IPTG for 6 hours at 30^0 C. The cells were pelleted out by centrifuging for 40 min at 9000 rpm at 4^0 C and then re-suspended in lysis buffer (20mM Sodium Acetate Buffer, 2mM TCEP, 0.1 % Triton, 1mM PMSF at pH 5.5). The cells were lysed by passing through French Press for 6 cycles at a pressure of 1200 psi. Lysed cells were centrifuged for 1 hr at 30000 rpm at 4^0 C. The protein present in the supernatant was passed through Cation Exchange Column and was eluted with 20 mM Sodium Acetate Buffer containing 1M NaCl and 2mM TCEP at pH 5.5. The eluted

protein was further passed through a C4 Reverse Phase Column (Higgins Analytical, Inc) and then lyophilized and stored at -20^0 C. Molecular mass was confirmed by mass spectrometry (Mr 7571.3 Da) and was > 99 % pure. Stability and reversibility of the protein were checked by Far UV CD temperature ramp.

2.4.2 Fluorescent-labeling of Engrailed Homeodomain

Engrailed homoeodomain with end cysteines was first labeled with the donor, Alexa 488 dye (Invitrogen), and then labeled with the acceptor, Alexa 594 (Invitrogen).

1. About 10mg of the unlabeled protein was added to 3 mL buffer (20mM Sodium Acetate, 100 mM NaCl and 2mM TCEP at pH 5.5) containing A488 dye. While preparing the above reaction mixture, first the dye was dissolved in 1 mL buffer and the protein was dissolved in 2 mL buffer and then the dye was slowly added in drops to the solution containing protein with constant stirring. The reaction mixture was then left overnight in the fridge for the labeling reaction.
2. About ~ 1 μL β-mercaptoethanol (Sigma Aldrich, > 99% pure) directly from the stock solution from the company was added to the above reaction mixture after the mono labeling reaction (the following day) in order to further make sure that both intra/inter molecular disulphide bonds are not formed before the purification procedure.
3. Mono-labeled protein was purified by passing the above reaction mixture through a C4 Reverse Phase Column. Typically, the fractions obtained from this purification would have unlabeled protein, protein labeled on both ends by Alexa 488 and protein labeled at only one end by Alexa 488. Fraction that contained

only the mono-labeled protein was identified by confirming the molecular mass from mass spectroscopy (M_r = 8295.1 Da). Excess Alexa 488 was completely removed by the following wash steps.

4. Fractions containing protein labeled with A488 obtained from the above purification was passed through a 3kDa centrifuge filter, by washing 3 times first with the 20mM Sodium Acetate buffer containing 100 mM NaCl and 2mM TCEP at pH 5.5 with 6 M Urea. This is to completely unfold the protein so that the dyes that non-specifically and non-covalently attached to protein would come out into solution and pass through the filter.
5. To remove urea, the supernatant obtained in the above wash step, containing the mono-labeled protein was washed three times with the 20mM Sodium Acetate buffer containing 100 mM NaCl and 2mM TCEP at pH 5.5 by using a centrifuge filter. Supernatant containing the mono-labeled protein was finally concentrated to ~ 2mL.
6. Alexa 594 dye was dissolved in 1 mL buffer (20mM Sodium Acetate, 100 mM NaCl and 2mM TCEP at pH 5.5) and then added slowly in drops to the solution containing the mono-labeled protein. The reaction mixture was then left overnight for the double-labeling reaction.
7. About ~ 1 μL β-mercaptoethanol (Sigma Aldrich, > 99% pure) directly from the stock solution from the company was added to the above reaction mixture after the double labeling reaction (the following day).
8. Double-labeled protein was purified by passing the above reaction mixture through a C4 Reverse Phase Column. Molecular mass corresponding to the

double-labeled protein was confirmed by mass spectroscopy and was M_r = 9153.6 Da. Fractions obtained from this purification would have unlabeled protein, protein labeled on both ends by the acceptor Alexa 594 and protein labeled at one end by the donor and another end by the acceptor. Only the fractions corresponding to the latter were further taken.

9. Fractions containing double-labeled protein obtained from the above purification was passed through a 3kDa centrifuge filter, by washing 3 times first with the 20mM Sodium Acetate buffer containing 100 mM NaCl and 2mM TCEP at pH 5.5 with 6 M Urea and then with the same buffer without urea. Supernatant containing the mono-labeled protein was finally concentrated to ~ 1mL.

Concentration of the protein was measured using the extinction coefficient of both the donor (A488, ε_{488nm} = 72,000 $M^{-1}cm^{-1}$) and the acceptor (A594, ε_{594nm} = 96,000 $M^{-1}cm^{-1}$). The average concentration of the double-labeled protein sample was 13.575 μM. This was then stored at -80^0 C as 20 μL aliquots.

2.5 Förster Resonance Energy Transfer Measurements

2.5.1 Förster Resonance Energy Transfer

Förster Resonance Energy Transfer (FRET) occurs when the emission spectrum of the donor overlaps with the absorption spectrum of the acceptor. This occurs because of long-range dipole-dipole interactions between the donor and the acceptor. The rate of energy transfer (k_T) is given by

$$k_T(r) = \left(1/\tau_D\right) \cdot \left(R_0/r\right)^6 \qquad (2.12)$$

in which τ_D is the lifetime of the donor in the absence of acceptor, R_0 is the Förster distance – which is the distance at which the energy transfer is 50 % and r is the distance between the donor and the acceptor.

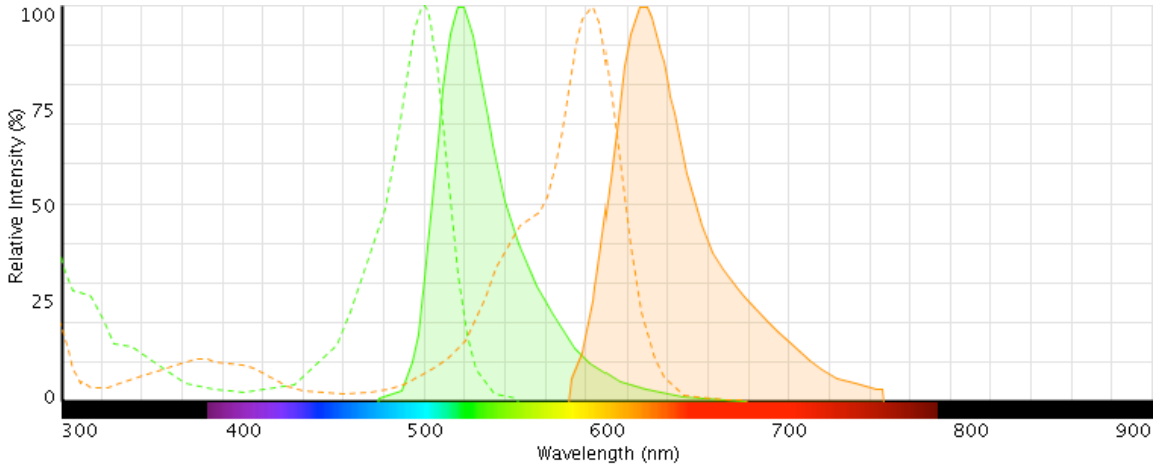


Figure 2-9. *Normalized absorption and emission spectrum of Alexa 488 and Alexa 594. Green dashed line and orange dashed line represent the absorption spectrum of Alexa 488 and Alexa 594 respectively. Green and orange filled areas represent the emission spectrum of Alexa 488 and Alexa 594 respectively.*

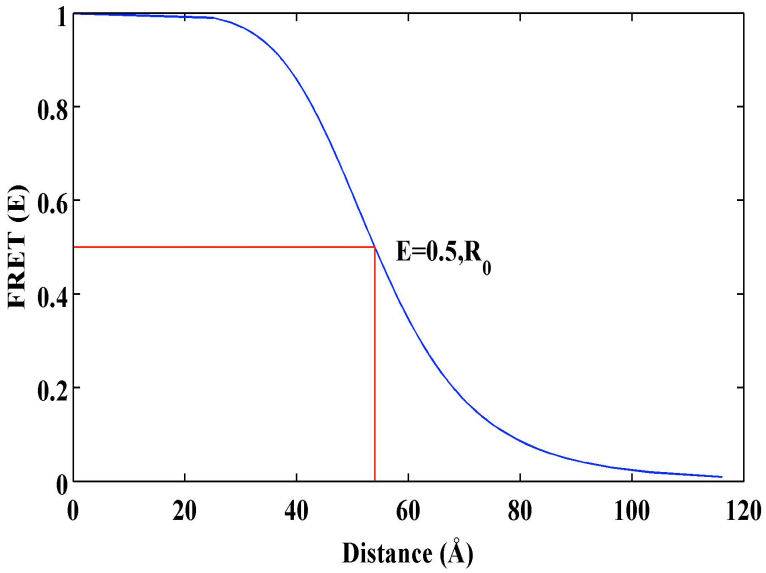


Figure 2-10. *Representative FRET Efficiency vs. Distance curve of Alexa 488/594 FRET pair. R_o for this FRET pair is 54 Å*

Energy transfer (E) is distance dependent and is given by

$$E = \left(R_0^6 \Big/ R_0^6 + r^6 \right) \tag{2.13}$$

Energy transfer depends on the overlap between donor and acceptor. It is dependent on the distance between the donor and acceptor, quantum yield of the donor and the relative orientation of donor and acceptor transition dipoles. Taking into account these factors, the rate of energy transfer (k_T) is given by the following expression.

$$k_T(r) = \left(Q_D \kappa^2 \Big/ \tau_D r^6 \right) \left(9000 \ln(10) \Big/ 128 \pi^5 N n^4 \right) J \tag{2.14}$$

where Q_D is the quantum yield of the donor in the absence of acceptor, τ_D is the lifetime of the donor in the absence of acceptor, κ is the orientation factor, N is Avogadro's number, n is the refractive index of the solvent and J is the overlap integral between the emission spectrum of the donor and the absorption spectrum of the acceptor.

The overlap integral (J) is calculated from the area under the curve and is given by

$$J = \left(\int F_D(\lambda) \varepsilon \lambda^4 \, d\lambda \Big/ \int F_D(\lambda) d\lambda \right) \tag{2.15}$$

where F_D is the emission of the donor and ε_A is the molar extinction coefficient of the acceptor.

The orientation factor (κ) is given by the following expression

$$\kappa = (\cos(\theta_T) - 3\cos(\theta_D)\cos(\theta_A))^2 \tag{2.16}$$

in which θ_T is the angle between donor emission transition dipole and acceptor absorption transition dipole, θ_D and θ_A are the angles between the donor and acceptor dipole planes and line or the vector joining those planes. When the rotational diffusion of

the molecule is too fast and randomized, the above value is approximated to 2/3 and is used to calculate R_0. For the dipoles that are more rigid, κ has to be calculated for that specific case.

Energy transfer is nothing but the ratio of transfer rate over total rate.

$$E = \left(k_T(r) \Big/ \left(1/\tau_D\right) + k_T(r) \right) \tag{2.17}$$

It can also be written in terms of fluorescence intensities and life times.

$$E = 1 - \left(F_{DA} \Big/ F_D \right) \tag{2.18}$$

$$E = 1 - \left(\tau_{DA} \Big/ \tau_D \right) \tag{2.19}$$

where F_{DA} is the emission of the donor in the presence of acceptor and F_D is the emission of the donor in the absence of acceptor. Similarly, τ_{DA} is the lifetime of the donor in the presence of acceptor and τ_D is the lifetime of the donor in the absence of acceptor. FRET efficiency also be expressed as function of number of photons hitting the donor and acceptor channel as follows:

$$E = \left(I_A \Big/ I_D + I_A \right) \tag{2.20}$$

where I_A and I_D are the number of photons emitted from of donor and acceptor respectively. FRET measurements are very common in protein folding, protein-protein, and protein-DNA interaction studies.

2.5.2 Sample Preparation for Bulk FRET measurements

Fluorescence measurements were done in a Fluorolog-3 spectrofluorimeter (Jobin Yvon) equipped with a thermostatted sample holder. For chemical unfolding experiments of fluorescent-labeled protein samples, bulk FRET was measured by exciting at 488 nm and the emission spectra were collected from 498 nm to 800 nm, every 10^0C from 5^0C to 35^0C. Concentration of the labeled protein for the fluorescence measurement was ~ 50 nM, obtained by diluting from 13.575 μM, 100 μL stock aliquot. First, a 500 nM stock was made from 13.575 μM labeled protein stock and then 500 nM stock was diluted ten times to obtain 50 nM required for the experiment. Samples were prepared in a quartz cuvette of 1 cm path length. Measurements were acquired using the following parameters: a slit width of 5 nm for excitation and emission slits and integration time of 0.25 s every nm.

2.5.3 Single Molecule FRET

Single molecule FRET measurements detect one single molecule in comparison to ensemble measurements in which one can observe only the bulk property of all the molecules. In other words, one can measure the distribution of conformation of biomolecules present at a particular experimental condition with single molecule measurements, in comparison to observing only the average value of the signal from these conformations with ensemble measurements. Detecting a single molecule above the background is no easy task. Typically, there is always Raman scattering contributed by the solvent as the solvent molecules are present at larger quantities when compared to the biomolecules of interest. In order to detect the FRET from single biomolecule above the

effect of scattering contributed by the solvent, the confocal volume should be as small as possible. In a typical folding measurement, ~ 1 fl confocal volume and concentrations as low as 25/50 pM are used to achieve this. There can also be a leak through of incident light, scattering from the molecules just above the confocal plane and stray lights. These are typically taken out as much as possible by suitable optics.

Single Molecule FRET Experimental Setup

Single molecule FRET measurements were carried out in a MicroTime 200 set up. (PicoQuant). Laser emitting at 488 nm was passed through to a fiber optics coupling unit, which attenuates the power of the laser drastically (~ 4x). Laser beam was aligned at its maximum intensity at this point. The excitation beam was then sent to the confocal microscope (Olympus – Inverted microscope). In an inverted confocal microscope set-up, the laser beam was then focused onto the sample (point illumination) by an objective (Numerical Aperture – 1.4, Oil Immersion - Type FF (Cargille)) present below the sample stage. The sample stage present above the objective was fitted with a temperature controller system. The sample chamber was covered by a plate and had a hole at the bottom in order to focus the light from the objective onto the sample. This plate helps maintaining the temperature of the sample inside the cuvette. A peltier was used to keep/monitor the temperature of the plate by an external controller. Water from the water bath was constantly circulated through the other side of the peltier in order to remove the excess heat generated by the peltier's hot plate and also to maintain the temperature at relatively similar temperatures of the water bath. For experiments, a calibration was made between the laser intensity (a.u.) at the sample stage to the power of the laser (μW).

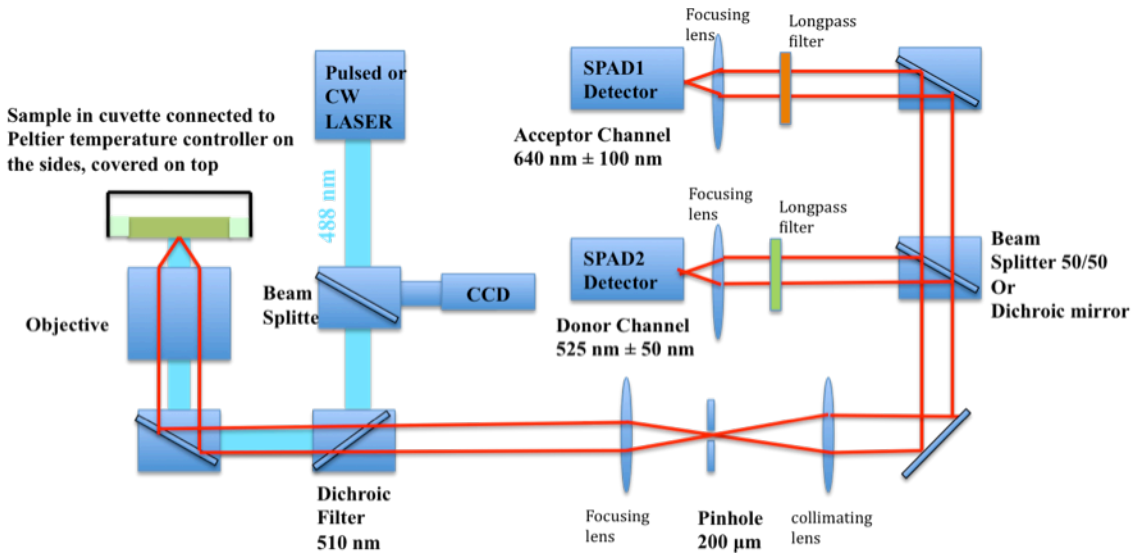


Figure 2-11. *Schematic of single molecule FRET experimental set up*

Laser power of ~ 150 µW was used chosen to use for our free diffusion single molecule FRET measurements to get good number of photons with the least photo damage. The measurement made in our case was a point measurement – measurement acquired at the same focal point. The focused laser beam on the sample was kept 20 µm above the cover slip surface into the buffer/liquid. A CCD camera was used to monitor the position of the surface of the cover slip and to help focusing the incident beam. The collected fluorescence light from the sample was passed through a dichroic filter (515 nm) and then focused on to the confocal pinhole aperture (200 µ), which helps cutting the light that are scattered from the sample and are not in focal plane, despite there was loss of some photons. Now, the beam was split into donor and acceptor detector channels through a dichroic (585 nm) mirror.

2.5.4 Sample Preparation for smFRET experiments

Cleaning of Cuvette: (Cuvette can hold up to 400 µL volume)

1. Unclean cuvettes glued to cover slips were put in chloroform solution in a beaker (1-2 days) to remove the cover slips.
2. Cuvettes were first rinsed thoroughly in water.
3. Cuvettes were then washed twice with acetone, methanol and water.
4. They were then sonicated for 30 minutes in 1M KOH solution.
5. After sonication, the cuvettes were rinsed with water again and left in acetone solution for 2 minutes.
6. Clean cuvettes were finally left for drying.

Cleaning Cover slips:

Cover slips were placed in the Teflon rack and dipped it in 10 % HF for 40 seconds followed by washing them in HPLC water. They were then left to dry.

Treatment of Buffers

1. Two spatula-full of charcoal (Carbon) was put into a falcon containing buffer solution (~ 30-50 mL).
2. Buffer solutions were kept shaking overnight.
3. Buffer solutions were taken using a syringe fitted with needle and then filtered using a 0.1 µm filter.
4. About 0.01% Tween 20 was added to the buffer solution.

charcoal treatment was basically done to remove invisible particulate matters that are present in the buffer solution as it would get adsorbed to charcoal.

Photoprotection cocktail (On the day of the measurement)

Trolox and Cysteamine specifically work for Alexa 488/594 donor – acceptor combination.

Stocks used were: Trolox (Sigma Aldrich) - 5.3 mg in 100 µL Methanol and Cysteamine (Sigma Aldrich) - 38.6 mg in 500 µL HPLC Water. pH of cysteamine was adjusted to the pH of the buffer.

Preparing Cuvettes:

1. A thin layer of glue (Norland UV glue) was applied on one side of the cuvette around the hole.
2. Cover slips were then carefully placed on the cuvette.
3. Cuvettes with cover slips were then glued using an UV lamp.

Preparation of the protein sample in the cuvette during the measurement:

1. A 20 µL double labeled protein aliquot (~ 13.575 µM) was taken and a series of stocks of 200 nM and 5 nM were made from that. Samples were prepared in a Lo Bind eppendorf.
2. From the 5 nM protein stock, 2 µL was taken to get a final concentration of 25 pM for a final volume of 400 µL in the cuvette for the measurement.

Sample in the cuvette:

Buffer (*)	392 µL
Trolox	2 µL
Cysteamine	4 µL
A488-Cys-Engrailed-Cys-A594	2 µL
Total Volume	400 µL

(*) – Buffer at the concentration of Urea required for the experiment

2.6 Singular Value Decomposition Analysis

SVD analysis helps obtaining the average experimental result and more information and details present in the experimental data. For example, SVD analysis[66] helps resolving the information present in a thermal unfolding spectra represented as a function of temperature (wavelength vs. temperature matrix). An experimental data, D (m x n – matrix) can be decomposed into three matrices as

$$Data, D = U \cdot S \cdot V^T \tag{2.21}$$

In these matrices, U and V^T are orthogonal and unitary and matrix, S has singular values of the data, D, on the diagonal and has values of zero elsewhere. Matrices, U, S and V have the sizes of (m x m), (m x n) and (n x n) respectively. When the matrix D is multiplied on the left by its transpose (D^T), it results in $D^T D = VS^2V^T$ and when the matrix is multiplied on the right by its transpose (D^T), it results in $DD^T = US^2U^T$. This is similar to writing the equation as $A = W\Lambda W^T$, in which A is a square and symmetric matrix, columns of W contain the eigenvectors of A and Λ is a diagonal matrix containing the corresponding eigenvalues. In simpler terms, the columns of matrix U are eigenvectors of DD^T, the columns of matrix V are of $D^T D$ and the square-root of the eigenvalues of the matrix $D^T D$ are the singular values of the diagonal matrix, S. This can be performed using the SVD subroutine that comes along with the MATLAB package. The singular values of the matrix S are ranked in the descending order and each value of S with respect to the first or the previous value of S gives the relative weight of that singular value. S when multiplied by columns of U gives the basis vector (spectra) and the corresponding amplitude vectors (temperature transition/decays) are given by the corresponding columns of V. Every column (or component) of U or V can give further

insights about experimental data that are not captured in the average experimental result and can be represented separately. Components that give rise to noise can be identified (lower singular values) and the experimental data can be reconstructed without the components that give rise to noise in the data.

SVD Rotation Analysis:

SVD rotation analysis is performed to improvise the results obtained from SVD analysis alone. In a SVD analysis, if the noise levels are comparable with the signal, they could come out as, as many numbers of significant components as the components contributing to the experimental signal. They may also be mixed in a SVD analysis. A noise component (uncorrelated amplitude component - V) can also come before the component that has a good signal (correlated amplitude component – V). SVD rotation analysis is an autocorrelation based transformation[66] (called rotation) procedure that is performed to improve the results obtained from SVD analysis in these cases. An autocorrelation function of V component for a given number of columns is given as follows. (It can be performed on either U or V depending upon what needs to be improvised)

$$C(V_i) = \sum_{j=1}^{n-1} \left(V_{j,i} \cdot V_{j+1,i} \right) \tag{2.22}$$

where $V_{j,i}$ is the *j*th element of *i*th column of V and n is the number of elements in each vector. A bad autocorrelation value (negative) for a particular V component (or column) in between good autocorrelation values (close to 1 or greater than 0.8) for the adjacent V

components can help identify the noisy components that are ranked higher in order according to singular values.

Rotation procedure introduces a linear transformation on selected columns (p) of V, which is denoted by V_k (where k=k1,k2,…,kp). The transformed vector V' is given by

$$V' = r_1 V_{k1} + r_2 V_{k2} + ... + r_p V_{kp} \tag{2.23}$$

The goal is to calculate or find the coefficients of transformation r_i (where i=1,2,.,p) of the transformed vectors V' such that the auto-correlations of the transformed vectors are optimized (maximum) with respect to autocorrelation performed without the transformation for the selected vectors. The entire mathematical derivation and solution can be found here[66].

2.7 Two State Analysis

2.7.1 Two State Analysis – Thermal Unfolding

In a two-state analysis, proteins are assumed to fold between only two distinct configurational states, native (N) and unfolded (U) states. It is approximated as first-order reversible reaction given by

$$N \rightleftharpoons U$$

Gibb's free energy is given by the following relation.

$$\begin{aligned} &\Delta G(T) = -R \cdot T \cdot \ln(K(T)) \\ &where \\ &K(T) = [U]/[N] \end{aligned} \tag{2.24}$$

The above relation can be reorganized to

$$K(T) = \exp\left(-\Delta G(T)/RT\right) \tag{2.25}$$

Equilibrium probability of a particular energy state (p_i) is the ratio of weight (w_i) corresponding to that particular energy state to sum of energy states. Partition function (Q) is defined as sum of energy states.

$$p_i = \left(w_i / \sum w_i\right) = \left(w_i / Q\right) \tag{2.26}$$

Folded state is taken as a reference and the energy corresponding to that state is set to zero. Thus, the weights corresponding to folded and unfolded states, partition function and probability of folded and unfolded states for a two-state scenario can be given by

$$w_N = 1; w_U = \exp\left(-\Delta G(T)/RT\right) = K(T) \Rightarrow Q = 1 + K(T) \tag{2.27}$$

$$p_N = \left(1/1 + K(T)\right); p_N = \left(K(T)/1 + K(T)\right) \tag{2.28}$$

A typical equilibrium-unfolding curve for a protein, with a single sigmoidal transition from a simple spectroscopic measurement, looks as in the figure 2-12. It has a pre-transition region, transition region and post-transition region. Equilibrium signal (S) obtained from a spectroscopic measurement for a two-state scenario is given by

$$< S >= p_N N + p_U U \tag{2.29}$$

where N and U are native and unfolded baselines and are given by

$$\begin{aligned} N &= S_{N0} + S_N (T - T_{ref}) \\ U &= S_{U0} + S_U (T - T_{ref}) \end{aligned} \tag{2.30}$$

where S_{N0} and S_{U0} are native and unfolded intercepts, S_N and S_U are native and unfolded slopes and T_{ref} can be any reference temperature that best fits the data for a two-state assumption. Folded and unfolded baselines represent the folded and unfolded signal values respectively for the two conformational states.

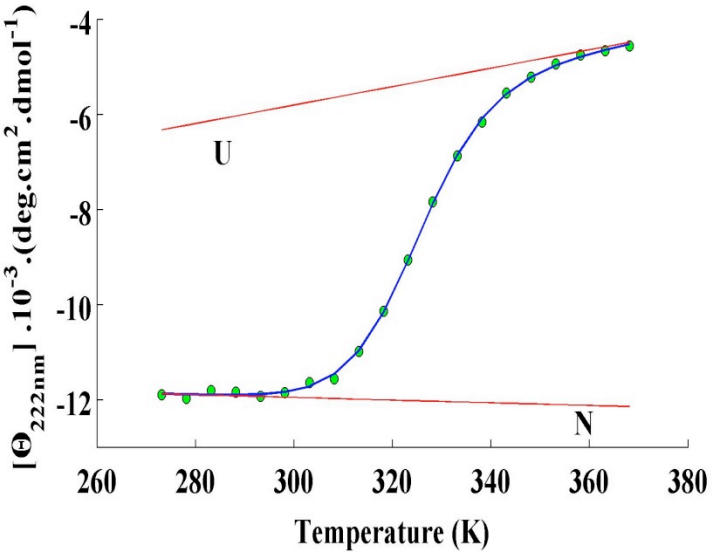


Figure 2-12. *Representative thermal unfolding curve. Two-state fit (blue line) of the experimental data (filled green circles) and baselines (red lines) from the two-state fit are shown in the figure.*

In the case of fitting an experimental DSC profile with a single peak to the above two-state equations, the experimental heat capacity ($C^p_{\exp}$) has to be described. $C^p_{\exp}$ is the sum of excess heat capacity (C^p_{ex}) and folded baseline (C^p_N). C^p_{ex} is the sum of intrinsic (C^p_{int}) and transition heat capacities (C^p_{trans}). They are given by

$$
\begin{aligned}
&C^p_N = N; C^p_U = U \\
&C^p_{\text{int}} = \Delta C_p \cdot p_U \\
&C^p_{trans} = \left(\Delta H^2_{cal} \cdot p_U - (\Delta H_{cal} \cdot p_U)^2\right) \Big/ RT^2 \\
&C^p_{ex} = C^p_{\text{int}} + C^p_{trans}; C^p_{\exp} = C^p_N + C^p_{\text{int}} + C^p_{trans} \\
&\textit{where} \\
&\Delta C_p = C^p_U - C^p_N
\end{aligned}
\qquad (2.31)
$$

ΔH_{cal} present in the above equations is the area between $C^p_{\exp}$ and $C^p_{\text{int}} + C^p_N$. For calorimetric data also, the native and unfolded heat capacity baselines, C^p_N and C^p_U, are given as in the equation 2.31. In order to fit experimental data to the above two-state equations, be it calorimetric data or from other spectroscopic probes, and obtain

thermodynamic parameters, free energy has to be expressed in terms of those parameters, which is given by Gibb's Helmholtz Free energy relation,

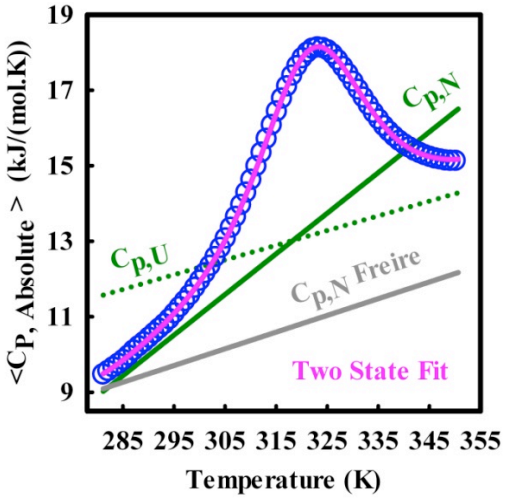


Figure 2-13. *Representative DSC profile. Two-state fit (pink) of the experimental data (blue circles) and the folded (green line) and unfolded baselines (green dot line) from the two-state fit are shown in the figure*

$$\Delta G(T) = \Delta H_m + \Delta C_p(T - T_m) + T \cdot (\Delta S_m + \Delta C_p \cdot \ln(T/T_m)) \qquad (2.32)$$

where ΔH_m is the vant-Hoff enthalpy (ΔH_{vH}) at T_m. At T_m, $\Delta G(T)$ =0, hence $\Delta S_m = \Delta H_m / T_m$. While performing a two-state fit, there are 6 fit parameters that needs to be calculated: ΔH_m, T_m, S_{N0}, S_N, S_{U0} and S_U. In this case, ΔC_p is allowed to vary with temperature. Typically, while fitting a DSC thermogram to a two-state model, ΔC_p is allowed to vary. ΔC_p can also be kept as constant while performing a two-state fit of an experimental data from a spectroscopic technique. In this case, there are 7 fit parameters. As ΔC_p obtained from this case is not as reliable as from obtaining from a DSC profile, it is typically assumed to be zero for the later case.

When vant-Hoff enthalpy (ΔH_{vH}) is equal to calorimetric enthalpy (ΔH_{cal}), folding scenario is considered two-state. If $\Delta H_{cal} > \Delta H_{vH}$ for a single-peaked DSC

transition, then it implies a more-than two-state transition and the corresponding DSC profile can be de-convoluted into n-states. As ΔC_p is defined as the difference between the folded and unfolded baselines in a DSC profile, if the baselines cross in the middle of the transition, it would show the system clearly deviates from a two-state scenario. Normally, a stringent $\Delta H_{cal} = \Delta H_{vH}$ criterion is imposed on the fit of experimental heat capacity data to see if the system deviates from two-state or not.

2.7.2 Two State Analysis – Chemical Denaturation

This section goes through the analysis of chemical unfolding curves by a two-state model, in which case the average protein signal measured as a function of denaturant is analyzed. The average signal is given as in the equation (2.29). In that, the unfolded and folded probabilities, p_U and p_N, are given as a function of denaturant (D) as:

$$p_N = \frac{1}{1+K(D)}; p_U = \frac{K(D)}{1+K(D)}, \text{ where } K(D) = \exp\left(-\Delta G/RT\right) \tag{2.33}$$

The folded and unfolded baselines as a function of denaturant:

$$\begin{aligned} N &= S_{N0} + S_N[D] \\ U &= S_{U0} + S_U[D] \end{aligned} \tag{2.34}$$

where S_{N0} and S_{U0} are native and unfolded intercepts, S_N and S_U are native and unfolded slopes. Equilibrium free energy (ΔG), that is the energy difference between the folded and unfolded states, varies linearly with the concentration of denaturant (D). A linear fit of ΔG as a function of [D] yields a slope (m-value) and an intercept ($\Delta G_{H_2 0}$) Intercept is the value of stability extrapolated at zero denaturant concentration. m-value is the change in the stability on the addition of denaturant and is related to solvent-accessible surface

area. Chemical denaturation midpoint (C_m) is the concentration of denaturant at which p_N=p_U=0.5 and ΔG=0.

$$\Delta G = \Delta G_{H_2 0} - m \cdot [D] \tag{2.35}$$

Analysis of individual chemical unfolding curves using the above two-state model yields 6 fit parameters, $\Delta G_{H_2 0}$, m, S_{N0}, S_N, S_{U0} and S_U.

2.7.3 Two State Kinetics – Thermal Unfolding

Experimental relaxations that are single exponential, are typically analyzed by a two-state model.

$$U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} N$$

where k_f and k_u are rate of forward and reverse reactions. Rate of change of any species is the difference between the forward and reverse reaction or in other words, the difference between the rate of formation and disappearance of that species. Rate of change of native and unfolded states (species) are given by:

$$\begin{aligned} \frac{d[U]}{dt} &= k_u[N] - k_f[U] \\ \frac{d[N]}{dt} &= k_f[U] - k_u[N] \end{aligned} \tag{2.36}$$

The observed relaxation rate k_{obs} is obtained by fitting the experimental decay to single exponential and it is the sum of forward and reverse rate constants, $k_{obs} = k_f + k_u$. At equilibrium, $k_f[U]_{eq} = k_u[N]_{eq}$. Thus, for two-state kinetics, equilibrium dissociation constant (K) is related to k_f and k_u and is given by

$$K = [N]_{eq} / [U]_{eq} = k_f / k_u \tag{2.37}$$

Then, Gibb's free energy at equilibrium becomes, $\Delta G_{eq} = -RT\ln(K) = -RT\ln\left(k_f / k_u\right)$. The equilibrium probabilities p_N and p_U are calculated as in the equation (2.28). In a two-state kinetic scheme, a transition state (TS) too higher in energy is assumed between the folded and unfolded states. Thus, the forward and reverse rate constants, k_f and k_u, are given by Arrhenius-like expression (Eyring's relation):

$$k_f = k_0 \cdot \exp\left(-\Delta G_f^{TS} / RT\right); k_u = k_0 \cdot \exp\left(-\Delta G_u^{TS} / RT\right) \tag{2.38}$$

where k_0 is the pre-factor or effective diffusion coefficient, D_{eff}, ΔG_f^{TS} is the free energy difference between the unfolded state and the transition state (TS) or for the forward reaction and ΔG_u^{TS} is the free energy difference between the folded state and the transition state (TS) or for the reverse reaction. Equilibrium free energy is the difference between free energies of the forward and reverse reaction, $\Delta G_{eq} = \Delta G_f^{TS} - \Delta G_u^{TS}$, where ΔG_f^{TS} and ΔG_u^{TS} for the forward and reverse reaction can be calculated according to the equation (2.32), in which case ΔC_p can be assumed to be zero or can be calculated. While performing a two-state kinetic analysis, there are 7 fit parameters: activation enthalpies (ΔH_f^{TS} and ΔH_u^{TS}), entropies (ΔS_f^{TS} and ΔS_u^{TS}), heat capacities ($\Delta C_{p,f}^{TS}$ and $\Delta C_{p,u}^{TS}$) for the forward and reverse reactions and a pre-exponential, k_0 or D_{eff}.

2.8 Calculation of Barrier Heights From Differential Scanning Calorimetry

Analysis of equilibrium thermal unfolding curves and the thermal unfolding kinetics by a two-state model as described before do not actually result in the calculation of barrier height for folding. Though, while analyzing single exponential relaxation rates

k_{obs} vs. T using a two-state model, ΔG_f^{TS} can be considered equivalent to the barrier height, the estimation of prefactor is difficult and as a consequence, the estimations of ΔG_f^{TS} as well. By any means, in the case of low-barrier scenario, the two-state definition cannot absorb the underlying folding mechanism in the other case. While using a statistical mechanical model, there's a continuous definition of reaction co-ordinate. It can account for both two-state and downhill scenarios and can also lead to an estimation of barrier height from the rate and calorimetry data. It is also relevant from the point of view of analyzing experimental DSC data, as heat capacity can be directly related to protein partition function.

A statistical mechanical model splits the order parameter into many states. A protein molecule interconvert between end-states ($native, I_N$ and $unfolded, I_U$) and passes through a series of states (i) in which any adjacent states are in equilibrium with each other.

$$I_{N,0} \leftrightarrow I_1 \leftrightarrow I_2 \leftrightarrow ...I_i ... \leftrightarrow I_{n-1} \leftrightarrow I_{U,n^{th}state}$$

If the weight of a particular state (energy state,i) is given by $\exp\left(-\Delta G_i / RT\right)$, then the partition function (Q) is given by the sum of the energy states as:

$$Q = \sum_i \exp\left(-\Delta G_i / RT\right) = \sum_i \exp\left(-\Delta H_i / RT\right)\exp\left(\Delta S_i / RT\right) \tag{2.39}$$

Now the probability of a particular energy state $\left(p_i\right)$ at a particular condition (experimental condition, say temperature, denaturant, pH) can be calculated such that $\sum_i p_i = 1$.

$$p_i = \frac{\exp\left(-\Delta G_i / RT\right)}{Q} \tag{2.40}$$

Having known the weight or the probability (p_i) of a particular state (i), any average property <A> (thermodynamic) of the system/protein molecule can be calculated.

$$\langle A \rangle = \sum_i A_i \times p_i = \frac{1}{Q} \sum_i A_i \times \exp\left(-\Delta G_i / RT\right) \tag{2.41}$$

From the free energy functional (G), defined in terms of a suitable order parameter (states, i), an apparent estimate of barrier height to folding and unfolding at different experimental conditions can be estimated. Two different statistical mechanical models were explored. Variable Barrier Model and Mean Field (One Dimensional Free Energy Surface Model) were used to fit the heat capacity unfolding curve from calorimetry. Barrier heights were obtained for various starting conditions for the fit. Robustness of the fit results was analyzed by a Bayesian approach and best result was chosen. The best result from the thermodynamics was further taken to kinetics by using a Kramer's like diffusion equation. The models, Bayesian Analysis, Rate analysis are discussed further in this Chapter.

2.8.1 Variable Barrier Model (VB Model)

Variable Barrier Model[67] is a one-dimensional free energy model derived based on Landau Theory of Critical Transitions. The theory describes the critical transition by a free energy functional (G) given as a function of an order parameter (x) by a Taylor series expansion. The expansion is truncated to the fourth power term. Such an expression produces free energy function between a single-welled minimum or two well-defined minima depending upon the sign of the quadratic term. In VB model, enthalpy is chosen

as an order parameter, that is, free energy (G) is represented as a function of enthalpy (H), G(H).

$$G(H) = G(H^2) + G(H^4) + ...$$
$$G_0(H) = -2\beta \cdot (H/\alpha)^2 + |\beta| \cdot (H/\alpha)^4 \quad (2.42)$$

where G_0 is the free energy surface at the characteristic temperature T_0, the temperature at which the two minimums in the free energy surface are equal, β and α are the two parameters that are expressed as coefficients of quadratic and the fourth power term. Case β < 0: there's going to be one minimum in the free energy surface defined as a function of enthalpy and α and β are parameters that best fit the heat capacity data. Case β > 0: there will be two minimums in the free energy surface at $H = \pm\alpha$ and a maximum at H=0 and β is the 'height of the free energy barrier' at the characteristic temperature, T_0. A fractional asymmetry factor (*f*) is introduced in the above definition to have asymmetric free energy surface on either side of H=0. If $\alpha_N + \alpha_P = \sum \alpha$, where α_N and α_P represent the α parameters for the negative and positive values of enthalpy, α_N and α_P are given by: $\alpha_N = \sum \alpha \cdot (f/2)$ and $\alpha_P = \sum \alpha \cdot ((2-f)/2)$. $\sum \alpha$ is an estimate of enthalpy and is the difference in the enthalpy between the two minimums at higher and lower temperatures. Now, there's a good definition of free energy surface. In order to fit the protein folding heat capacity data using this model, that is, to obtain $C_{p,\exp}$, first and second order moment need to be calculated according to the following equation.

$$\langle H^n \rangle = \int H^n \cdot p(H/T).dH \quad (2.43)$$

Q and p(H/T) can be calculated according to equations (2.50) and (2.51) (this is just a continuous expression of the same). Using <H> and <H^2>, one can calculate $C_{p,\exp}$ using

equations (2.31). Thus, four parameters: $T_0, \sum\alpha, \beta(T_0)$ and f are used to fit the experimental heat capacity data using the above model.

2.8.2 One Dimensional Free Energy Surface Model (MF Model)

Mean Field Model[68] uses nativeness (n) as a reaction coordinate, where nativeness is the probability of finding a particular residue in a native-like conformation, which is a modified Zwanzig parameter[69]. Here, nativeness (n) represents the progress of a reaction towards the folded side as a continuous reaction co-ordinate. As conformation or configurational space in terms of probability (n), entropy cost associated with it can be calculated by taking into account all possible native (n) and non-native conformational probabilities (1-n) based on Gibb's entropy formulation: $\Delta S = -R\sum p_i \ln(p_i)$. Such a formulation as a function of nativeness is given as:

$$\begin{aligned} &\Delta S(n) = N\cdot\left[-R\cdot\left(n\cdot\ln(n)+(1-n)\cdot\ln(1-n)\right)+n\cdot\Delta S_{res}^{n=1}+(1-n)\cdot\Delta S_{res}^{n=0}\right] \\ &\textit{where} \\ &\Delta S_{res}^{n=1}=0;\Delta S_{res}^{n=0}=\Delta S_{res} \end{aligned} \tag{2.44}$$

where N is the number of residues present in the protein and ΔS_{res} is the entropy cost associated with changing a particular residue from native-like conformation to complete non-native/unfolded conformation. An estimate of 16.5 $J\cdot mol^{-1}\cdot K^{-1}\cdot res^{-1}$ is taken for ΔS_{res}. Enthalpy ($\Delta H(n)$, stabilization energy), heat capacity ($\Delta C_p(n)$) and free energy ($\Delta G(n)$) as a function of nativeness are given as:

$$\Delta H(n) = N \cdot \Delta H_{res} \cdot [1 + \frac{\exp(k_{\Delta H} n - 1)}{(1 - \exp(k_{\Delta H}))}]$$

Similarly,

$$\Delta C_p(n) = N \cdot \Delta C_{p,res} \cdot [1 + \frac{\exp(k_{\Delta C_p} n - 1)}{(1 - \exp(k_{\Delta C_p}))}] \tag{2.45}$$

$$\Delta G(n) = \Delta H(n) - T \cdot \Delta S(n)$$

where ΔH_{res} and $\Delta C_{p,res}$ are the stabilization energy and heat capacity per residue; $k_{\Delta H}$ and $k_{\Delta C_p}$ are model parameters that best describe the functional vs. n. Incorporating temperature changes in the above functions:

$$\Delta S(T,n) = \Delta S(n) + \Delta C_p(n) \cdot \ln(T / 385)$$
$$\Delta H(T,n) = \Delta H(n) + \Delta C_p(n) \cdot (T - T_{ref}) \tag{2.46}$$
$$\Delta G(T,n) = \Delta H(T,n) - T \cdot \Delta S(T,n)$$

Now, having the free energy surface and other thermodynamic parameters represented as a function of both nativeness and temperature, one can calculate $C_{p,\exp}$ using equations (2.31). Now, one can fit the experimental heat capacity data using this simple model and it uses only two parameters for the fit: ΔH_{res} and $k_{\Delta H}$. Empirical estimates of 4.3 and 50 $J \cdot mol^{-1} \cdot K^{-1} \cdot res^{-1}$ for $k_{\Delta C_p}$ and $\Delta C_{p,res}$ respectively are typically used for the fitting. A value of 58 $J \cdot mol^{-1} \cdot K^{-1} \cdot res^{-1}$ for $\Delta C_{p,res}$ can also be used, which is based on the data reported in the literature. A modified version of this model[70], which incorporates local and non-local contributions to enthalpy, has been used lately. In this case, enthalpy function is represented as a sum of local and non-local contributions.

$$\Delta H(n) = \Delta H_{local}(n) + \Delta H_{non-local}(n) \tag{2.47}$$

Enthalpy equation in (2.41) is split into two.

$$\Delta H_{local}(n) = N \cdot \Delta H_{res}^{local} \cdot [1 + \frac{\exp(k_{\Delta H - local} n - 1)}{1 - \exp(k_{\Delta H - local})}] \tag{2.48}$$

$$\Delta H_{non-local}(n) = N \cdot \Delta H_{res}^{non-local} \cdot [1 + \frac{\exp(k_{\Delta H-non-local} n - 1)}{1 - \exp(k_{\Delta H-non-local})}] \quad (2.49)$$

in which, the values of $k_{\Delta H-local}$ and $k_{\Delta H-non-local}$ are parameterized to be -1.25 and 3.75 respectively. In this new version of the model, again there are only two parameters that need to be fit, ΔH_{res}^{local} and $\Delta H_{res}^{non-local}$. Barrier height can be obtained from the free energy surface. Melting temperature is the temperature at which $\Delta\Delta G_{eq}$=0. $\Delta\Delta G_{eq}$ is the difference in the free energy between the two minimums in the free energy surface ($\Delta G(T,n)$) at any given temperature. Calorimetry profiles were fit using the modified version of mean field model. Single molecule data were still fit with the old version of the model.

2.8.3 Bayesian Analysis

Analysis of DSC profile by multiple free energy surface models for various starting assumptions/criterions can yield different parameters. This leads to questions of what is the best model to use, what is the best fit and what are the parameters that would better describe the experimental data. Thus, a Bayesian analysis is used to assess the quality of the fits and rank[70] them based on the results obtained from the fits.

In this particular case, sum of least squares (SLS) and native baseline slope were chosen to assess the quality of the fits. Apart from taking into account the fit goodness (SLS), how far the native baseline slope deviates from the reference slope could provide meaningful assessment of the fit results from calorimetry.

Using Bayesian analysis, Probability (P_h) of the hypothesis (h) – native baseline slope (b) is calculated based on the error (SLS) obtained from the fit of the experimental

data (D) profile to these models. In other words, fits are according to P_h and the best fit and parameters are chosen. Also, we can have an estimation of an average barrier height to folding (β_h).

$$P_h = P(h \mid D) = \frac{P(h) \cdot P(D \mid h)}{P(D)} \tag{2.50}$$

$$\begin{aligned} P(h) &= \frac{1}{\sigma_b \sqrt{2\pi}} \cdot \exp\left(-\frac{b - b_0}{2\sigma_b^2} \right) \\ P(D \mid h) &= \frac{1}{Z} \cdot \exp\left(-N_{eff} \cdot \frac{SLS / N}{2\sigma^2} \right) \end{aligned} \tag{2.51}$$

P(D)=constant for a given set of experimental data

$$P_h = P(h \mid D) = \alpha \cdot \prod\nolimits_h = \alpha \cdot \exp\left(-\gamma \cdot SLS - \frac{(b/b_0 - 1)}{2(\sigma_b / b_0)^2} \right) \tag{2.52}$$

where α and γ are constants, b is the native baseline slope obtained from the fit and b_0 and σ_b are the slope and standard deviation for the reference baseline. Reference slope typically chosen for the analysis of thermal unfolding curves from calorimetry is the Freire's native heat capacity baseline slope and the corresponding values of b_0 and σ_b are 0.0067 $M_r J \cdot K^{-2} \cdot mol^{-1}$ and 0.0013 $M_r J \cdot K^{-2} \cdot mol^{-1}$, where M_r is the molecular weight of the protein in $g \cdot mol^{-1}$. The values of SLS and b are already known from the fit results. P(D) and other constants from P(D|h) and P(h) before the exponential terms are embedded in α.Right γ has to be found for performing the analysis. As P_h is proportional to $\prod_h$, fit result that yields the highest value for the normalized values of $\prod_h$ is the best-fit result from this analysis.

An entropy (Shannon entropy) is calculated from the normalized $\prod_h$ values (equivalent to calculating entropy from probability) in order to find the right value of for γ that can be used in the above probability expression. Shannon entropy is defined as $S_{SH} = -\sum_h \prod_{h,Norm} \cdot \ln\left(\prod_{h,Norm}\right)$ and is calculated for a given set of γ values. Shannon entropy is normalized as, $f_{SH} = S_{SH}(\gamma)/S_{SH}(\gamma = 0)$. Now, f_{SH} is between zero and 1. Rather, f_{SH} is normalized with respect to the f_{SH} value for the lowest value of γ. For lower γ values, the probability will be biased towards the slope values ($f_{SH} \to 1$). For higher γ values, SLS values dominate the probability ($f_{SH} \to 0$). To circumvent this problem, γ corresponding to $f_{SH} = 0.5$ is chosen. Now, $\prod_{h,Norm}$ is calculated for this value of γ and fits are ranked.

This Bayesian analysis approach helps evaluating the quality of the fits obtained from DSC data using different robust one-dimensional free energy surface models for various assumptions and point out the best-fit results.

Average value of barrier height to folding is calculated using this probability: $\langle \beta_h \rangle = \sum_{i=1}^{no.of.fits} \beta_h(i) \cdot \prod_{h,Norm}(i)$. The corresponding standard deviation associated with it is calculated as: $\sigma_\beta = \sqrt{\langle \beta^2 \rangle - \langle \beta \rangle^2}$. In this equation, the second moment $\langle \beta^2 \rangle$, is calculated using this expression: $\langle \beta_h \rangle = \sum_{i=1}^{no.of.fits} \beta_h^2(i) \cdot \prod_{h,Norm}(i)$.

2.9 Simulation of Temperature-Jump Decays

Experimental decays are simulated using Mean Field (MF) Model, in which the kinetics is described as diffusive by a Kramer-like treatment. The effective diffusion coefficient is given by,

$$D_{eff}(T) = k_0 \cdot \exp\left(-\frac{N \cdot E_{a,res}}{R \cdot T}\right) \tag{2.53}$$

where k_0 is a pre-exponential factor, N is the number of residues and $E_{a,res}$ is the activation enthalpy per residue. Thus, equilibrium and kinetics can be described using four parameters: 2 for equilibrium (ΔH_{res}^{local} and $\Delta H_{res}^{non-local}$) and 2 for kinetics ($k_0$ and $E_{a,res}$). Experimental temperature jump decay can be completely simulated/fit using these four parameters.

A rate-matrix method[72] for diffusive kinetics is employed while using MF Model to simulate the experimental T-Jump relaxation. In MF model, the order parameter used is nativeness (n). From this, we have the probability distribution as a function of n at every initial and final temperature of the T-Jump. We also have effective diffusion defined for every value of temperature. Base on this, a rate matrix is defined for every possible values of n as follows:

$$Rate(n \times n) = \begin{pmatrix} Rate(1,1) & Rate(1,2) & 0 & 0 & 0 \\ Rate(2,1) & Rate(2,2) & ... & 0 & 0 \\ 0 & Rate(3,2) & ... & Rate(i-1,i) & 0 \\ 0 & & ... & Rate(i,i) & Rate(i,i+1) \\ 0 & 0 & 0 & Rate(i+1,i) & Rate(n,n) \end{pmatrix} \tag{2.54}$$

where

$$Rate(1,1) = -Rate(2,1); Rate(n,n) = -Rate(n-1,n)$$
$$Rate(i,i) = -Rate(i+1,i) - Rate(i-1,i)$$
$$Rate(i,i+1) = (D_{mat}(T) \cdot p(i,T) / p(i+1,T) + D_{mat}(i+1)) / 2$$
$$Rate(i+1,i) = (D_{mat}(T) + D_{mat}(i+1) \cdot p(i+1,T) / p(i,T)) / 2$$
$$D_{mat}(T) = D_{eff}(T) / dn^2$$

where 'i' represents i^{th} state in the nativeness space, p(T) is the probability distribution vs. nativeness at final temperature of the T-jump and dn is the nativeness interval between two adjacent states.

Rate matrix is solved by eigenvalue analysis (using MATLAB). Eigenvalues and vectors obtained are used to calculate the survival probabilities at every time (t) point at a particular final temperature of T-Jump: $S(t) = \sum_n A \cdot \exp(\Lambda \cdot t)$, where n represents that S(t) is calculated over all possible values of nativeness and Λ represents the eigenvalues. Survival probability, together with a function that describes the average signal (Sig(n)) at a particular wavelength vs. n, experimental decays (<Sig(t)>) can be simulated as follows: $\langle Sig(t) \rangle = \sum_n Sig(n) \cdot S(t)$. Infrared experimental decays at two frequencies were fit this way. Entire spectral decay can also be simulated by this method. Fluorescence T-Jump spectral decays were fit this way. Results are shown in Chapter 4.

2.10 Analysis of single molecule FRET Trajectories

Single molecule FRET can be used to observe all possible conformational states of biomolecules. In the case of protein folding, it gives the 'real probability distribution' of a protein at every experimental condition.

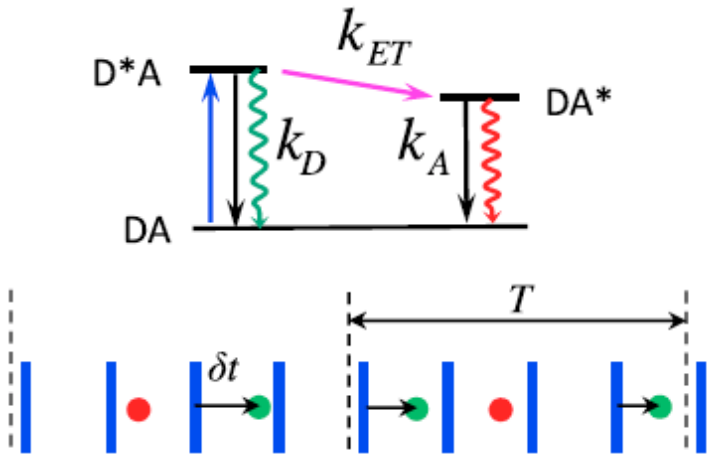


Figure 2-14. Schematic of different time processes in FRET (upper) and smFRET trajectory (lower). Figure adapted from the reference [73]

A typical set of photon bursts from a free diffusion single molecule FRET experiment is shown in the Figure 2-15. A burst is a series of photons in quick succession. In a typical free diffusion smFRET measurement, arrival time of every single photon (red circle – acceptor photon; green circle – donor photon in the figure 2-14) is recorded. Along with that, the delay time δt for both the donor and acceptor photon, that is, the difference in time between the laser pulse (blue vertical line in the figure 2-14) and the photon (red/green circle in the figure 2-14) is also recorded. FRET value is typically calculated as the ratio of number of acceptor photons recorded in the acceptor channel to total number of photons recorded in both the channels (donor and acceptor) within the same time interval called time bin, T. Binning is a procedure by which a time series data are divided into equally spaced time bins (intervals) and the number of photons falling within each interval are allocated to that time bin. Analysis of smFRET trajectories (time-series data) require i) selection of photon bursts and then ii) extracting thermodynamic/kinetic information from the photon arrival times data using appropriate models. In this section, we describe a clustering approach to better select for the photon

bursts and implementation of mean field model (1DFES Model) by maximum likelihood parameter optimization to analyze the single molecule experimental data.

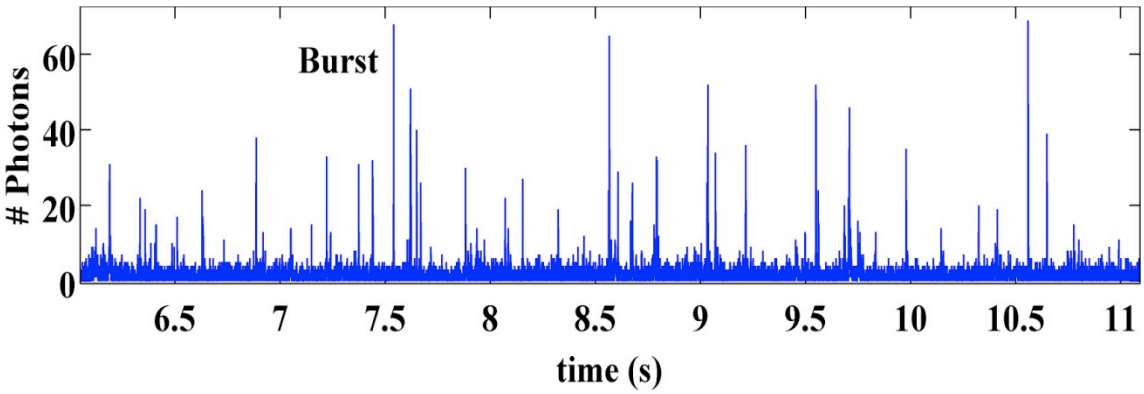


Figure 2-15: *smFRET trajectory*

2.10.1 K-means Clustering

K-means is a data clustering procedure that has various applications, in which, k represents the number of clusters present in the data. K-means clustering algorithm aims to minimize the distance between the data points (x) belonging to a particular cluster (k_i) to the point that defines the center of the cluster (μ_i) as follows:

$$\min \sum_{i=1}^{k} \sum_{x_j \in k_i} \left|\left| x_j - \mu_i \right|\right|^2 \tag{2.55}$$

First, number of clusters present in a given data set has to be given or assumed. There are procedures that help determine the number of clusters, yet in a typical k-means clustering algorithm, this is given as a user input. Once the number of clusters is defined, center (centroid) of each cluster is initialized. Now, each data point is assigned to a particular cluster based on how close the data point is to the cluster center. The above equation is calculated and the procedure is repeated until it converges or reaches a minimum. This procedure is applied in extracting the photons corresponding to photon bursts in smFRET trajectory data containing photon arrival time information. In a binning

analysis, time trajectories are divided into equal time bins (T). The events greater than a certain threshold (> certain number of photons) are selected as bursts from the time-series data uniformly analyzed using the same time bin interval. But, while analyzing the time-series data (photons vs. time) using k-means clustering procedure, it selects every burst at the appropriate time interval and retains the arrival time data. Each cluster represents a burst and every particular cluster has 'n' number of photons and has a unique time-length. Rather, every particular burst/cluster has a different time-length.

2.10.2 One Dimensional Free Energy Model by Maximum Likelihood Method

Maximum likelihood method is a parameter optimization method, that is, given a distribution or a model that describe the data, this method finds the best set of parameters that describe the data based on that distribution/model. For example, if we have a Gaussian distribution, the method would estimate the parameters, mean and the variance. For a given set of observation/data ($y = y_1, y_2, ..., y_n$) that are independent of each other described using a probability density function with the parameters ($\theta = \theta_1, \theta_2 ..., \theta_n$), the overall probability density function can be given as a multiple of each other: $p_1(y_1 / \theta).p_2(y_2 / \theta)...p_n(y_n / \theta)$. Each conditional probability given in this equation is called the likelihood function (L). The parameters that best fit the data are the ones that maximize the likelihood function. Typically, log-likelihood (log(L)) is calculated. In that case, log-likelihood is given by

$$\log(L) = \sum_{i=1}^{n} \log(L_i)$$
$$\textit{where} \qquad (2.56)$$
$$L_i = p_i = p_i(y_i / \theta)$$

In our case, we implement a likelihood method to analyze the photon arrival times to fit the experimental data using MF model. We want to get best fit parameters: 2 for equilibrium (ΔH_{res} and $k_{\Delta H}$.), $D_{mat}(T)$ diffusion parameter for kinetics (D) and 2 for the FRET function represented as a function of nativeness(n) $FRET_0$ and $\Delta FRET$ (see below), for this statistical mechanical model, that would best describe the entire single molecule data. If we can fit the data using this model, this would also help estimate the barrier height to folding β_F from the free energy surface.

FRET occurs at a faster time-scale in comparison to inter-photon time scales. Photon arrival is independent of each other and thus, the statistics can be described by a Poisson distribution. A likelihood function is derived based on this for every burst. Likelihood is calculated for every photon present in a particular burst and likelihood for the entire burst is calculated as a multiple of likelihoods for all the photons in that burst. As two bursts (events) present in a time trajectory are independent of each other and once we have the likelihood function for a single burst, it can be calculated for the entire time-series data as a multiple of each other.

Likelihood function[74] for a single burst is given by:

$$L = 1^T . \prod_{k=2}^{No.of.Photons/burst} (F(c_k)e^{K\cdot\tau_k}) \cdot F(c_1) \cdot p_{eq} \qquad (2.57)$$

where 1^T is a unit vector, p_{eq} is a vector containing equilibrium probabilities, $F(c_k)$ is a photon color matrix(In the case of two colors (donor and acceptor), if its E for donor, it is going to I-E for acceptor, where I is the unity matrix), K is the rate matrix and τ_k is the inter-photon time between the k^{th} photon and the one before that. This multiplication is done from the second photon as the extreme terms $F(c_1)$ and p_{eq} represent the photon

color and the equilibrium probabilities for the first detected photon and the equilibrium probability of the system at the beginning of the burst. Typically, while evaluating this above equation, the rate matrix is diagonalized and hence can be solved by eigenvalue analysis. The above equation then becomes,

$$L = p_0^T . \prod_{k=2}^{No.of.Photons/burst} (\Phi(c_k) e^{\Lambda \cdot \tau_k}) \cdot \Phi(c_1) \cdot p_0 \qquad (2.58)$$

Rate matrix for statistical mechanical model (Mean Field Model) can be written as described in the section 2.54. Λ is an eigenvalue matrix and can be obtained by solving the rate matrix. A FRET function is defined for every value of the order parameter, nativeness (n) as $FRET(n) = FRET_0 + n \cdot \Delta FRET$. From the eigenvector and the FRET function, we can calculate $\Phi(c_k)$ for the acceptor. From this, we can calculate $\Phi(c_k)$ for the donor as $I - \Phi_k$. $\Phi(c_1)$ is also calculated similarly for the first photon depending upon whether it is a donor or acceptor. The p_0 vector holds just only one non-zero value of 1 corresponding to the zero eigenvalue (slowest eigenvalue) corresponding to equilibrium starting point. Now, we can calculate the likelihood for one particular burst/cluster and hence the log-likelihood. This is repeated for all the burst and the over all log-likelihood is calculated as the sum of likelihoods for individual bursts. The set of parameters that maximizes this log-likelihood are the best-fit parameters. In the case of evaluating negative log-likelihood for every burst, the sum of negative log-likelihood for all the burst is calculated and the minimum criterion is chosen to get the best-fit parameters. Now that we have best fit parameters, we will have the free energy surface and this would help elaborating the folding mechanism of the protein as conventional two-state or downhill folding mechanism.

*Integration of maximum likehlihood with one dimensional free energy surface model was done by one of the colleagues in our lab. It was based on synthetically creating the burst and then analyzing them. In the current research, this was applied and extended to the experimental data.

Chapter 3: Equilibrium thermal unfolding of Engrailed Homeodomain by Multiple Spectroscopic Probes

3.1 Abstract

Engrailed homeodomain, a small α-helical domain that folds in microseconds, can be a likely downhill folding candidate. In order to test this, we study the thermal unfolding of EnHD by four spectroscopic probes together with that of differential scanning calorimetry. a) Calorimetry revealing a broad unfolding and crossing of folded and unfolded baselines, b) spread of melting temperature between different spectroscopic probes, c) wavelength-dependent unfolding by infrared and d) fluorescence showing complexities in the unfolding in the form of unveiling three different behaviors such as contact quenching, spectral shift and FRET transfer between aromatic amino acid residues were the highlights of the results obtained from these measurements. In order to uniformly describe all these complex unfolding behaviors observed, a global analysis of all the thermal unfolding experiments using MF model was done. This led to an estimation of barrier height to folding for EnHD near the characteristic temperature to be ~ 0.47 RT (<2RT) and thus falling within the 'downhill folding' regime.

3.2 Introduction

Many folding[49-57] and binding experiments[46,47,60] have been performed on engrailed homeodomain. Early calorimetry study and fluorescence temperature jump studies as a function of temperature on EnHD were described by a conventional two-state mechanism. This domain was one of the first reported microsecond-folding domains or single domain proteins. Further fluorescence T-Jump measurements as a function of temperature and fluorescence T-Jump measurements as a function of denaturant concentration at room temperature reported the presence of an additional faster phase and interpreted the results with a conventional three-state mechanism. Some molecular dynamic simulations results[75-83] supported the view of existence of an intermediate.

Keeping these results on one-hand and going by the fact EnHD is a microsecond folder and based on theoretical conclusions on the size-scaling of barrier heights validated by experimental results, it could be argued that it is highly likely that the domain could fold with a smaller barrier. All these studies mentioned before on this domain did not estimate the barrier height for folding and inherently assumed the presence of large barrier(s). If this domain folds with a small barrier or exert a downhill folding mechanism, then the nature of intermediate(s) must be argued in a different manner. This is because an intermediate separated by large barriers from the extreme states will depopulate when the denaturant stress is increased. On the contrary, the intermediate(s) separated by marginal barriers (< 2 RT) with respect to the native and unfolded states, would continuously unfold or keep on becoming unstructured on increasing the concentration of denaturant. In contrast to the some MD simulation results that argued for an accumulation of an intermediate, another MD simulation[84] studies that

made a big impact by simulating a series of fast-folding proteins reported facing problems, while trying to simulate this domain using the same crystal structure that was used to simulate before. These later arguments suggest revisiting the folding studies on this domain from a different perspective, yet conforming to the earlier experimental results. In order to support the later view on the folding studies of this protein, apart from performing similar experiment more experiments in terms of additional spectroscopic probes will need to be done. Even a multiple probe thermal unfolding measurements and a more detailed analysis of the results obtained from these measurements would provide sufficient information to support the downhill nature proposed for the folding mechanism of this protein.

3.3 Results

3.3.1 Differential Scanning Calorimetry (DSC), Far UV Circular Dichroism (fCD) and Near UV Circular Dichroism (nCD)

DSC: DSC thermo gram of EnHD revealed a broad transition with a clear peak around 323 K. A look at the experimental heat capacity data at low temperatures showed a steep temperature dependence of the heat capacity, whereas at high temperatures the changes in heat capacity with temperature become much smaller.

fCD: fCD spectrum at native conditions (298 K) obtained was very typical of α-helical proteins with minimums near 222 nm and 208 nm. The minimum at 222 nm was less well defined than expected for α-helical proteins. This was due to the presence of a set of aromatic residues in the core of EnHD contributing to the CD signal in that region. This would lead to an underestimation of the percentage secondary structural content

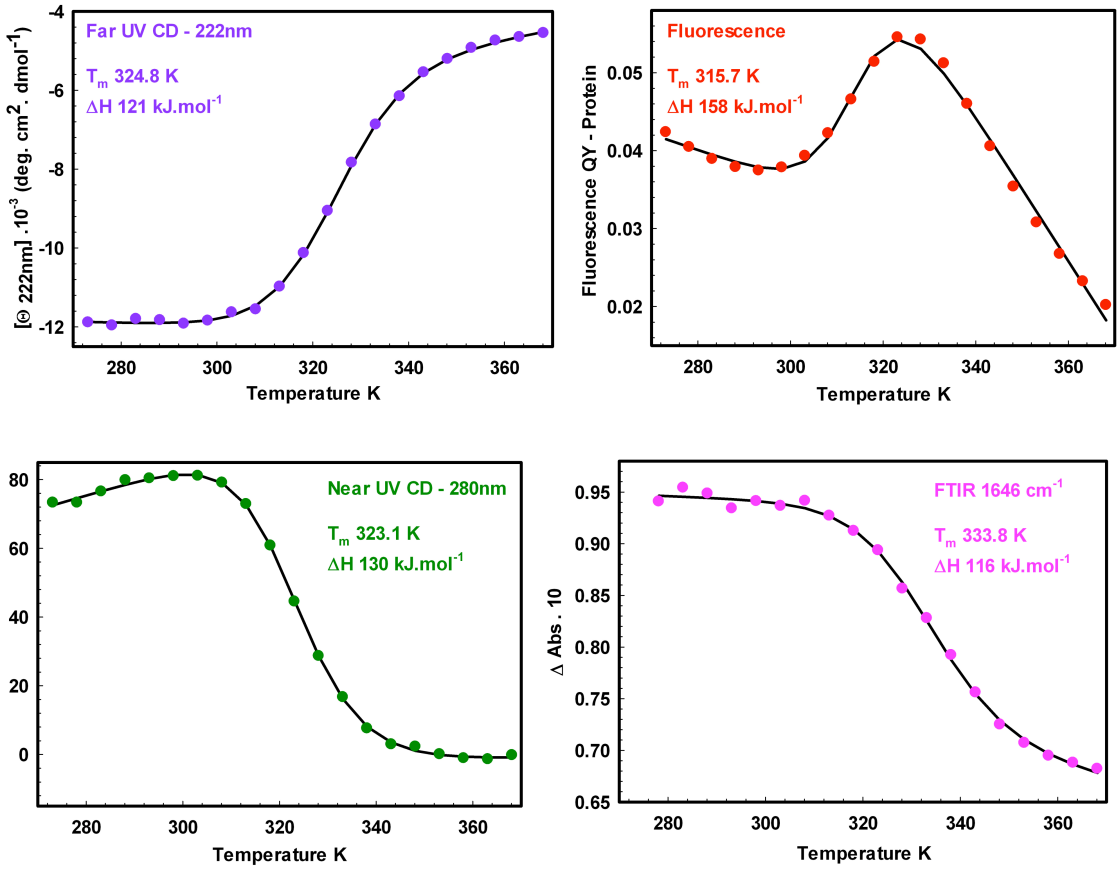


Figure 3-1. *Experimental unfolding curves from four spectroscopic probes (filled circles - experimental data) and the corresponding two-state fits (black line). Thermodynamic parameters obtained from the two-state analysis are indicated.*

calculated derived from the CD signal around this region and an estimation of percentage helical structure calculated using the ellipticity value at 208 nm[85] yielded a helical content of approximately 46% which was lesser than the percentage helical content calculated using DSSP algorithm from the crystal structure (~ 70%). The thermal unfolding transitions at 208 and 222 nm showed clear sigmoidal curves. The unfolding curve at 208 nm showed an increase in the CD signal (amounting to increase in the

secondary structural content with temperature) with temperature at the lowest temperatures measured.

nCD: nCD spectrum was very complicated to de-convolute as the nCD signal had contributions from both the secondary and tertiary structural contents of the protein. Unfolding curve at 280 nm showed a cooperative-like transition.

3.3.2 Fourier Transform Infrared Spectroscopy (FTIR)

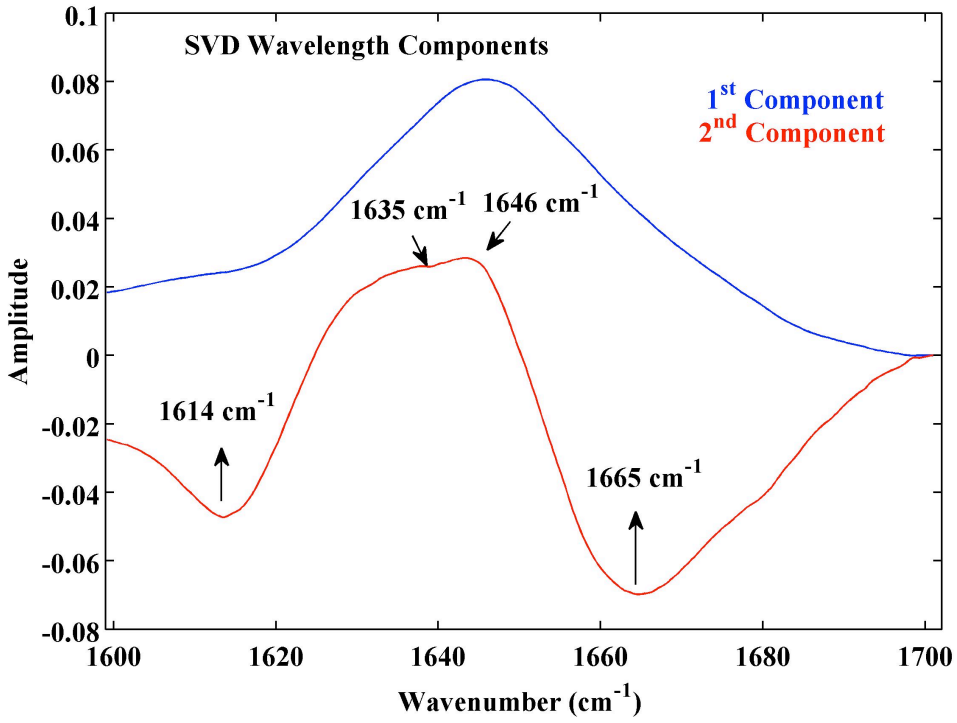


Figure 3-2. *SVD analysis of FTIR unfolding spectra. First two wavelength components (U) are shown in blue and red respectively. Also, the wave numbers corresponding to alpha helical signals (1646 cm^{-1} & 1636 cm^{-1}), random coil or unfolded signal (1665 cm^{-1}) and amino-acid side chain signal (1614 cm^{-1}) are indicated.*

FTIR spectrum of the EnHD was acquired in the Amide I region (1600 cm^{-1} - 1700 cm^{-1}) and thermal melting curves were obtained. The spectrum at the lowest

temperature measured, showed a peak around 1646 cm^{-1}, characteristic of α-helical proteins. Absorbance at 1646 cm^{-1} decreased with temperature, indicative of a decrease in the secondary structural content with a corresponding increase in the unfolded signal value around 1665 cm^{-1}. The second wavelength component from a simple SVD analysis of the unfolding spectra revealed a shoulder around 1635 cm^{-1} near the peak value of 1646 cm^{-1} and also a very small peak near 1614 cm^{-1}. It was very well reported in the literature that the FT-IR signal of alpha helical peptides/proteins could resolve into two peaks, one corresponding the solvent exposed part and another for the buried helical[86] content. Also, certain amino-acid side chains[64] of tyrosine, arginine, glutamine and asparagine could contribute signal near 1614 cm^{-1}. A less co-operative thermal unfolding for the solvent exposed α-helical signal at 1635 cm^{-1} was followed by a more co-operative transition at 1646cm^{-1}, which was further followed by an increase in the signal around 1614 cm^{-1}.

3.3.3 Steady-State Fluorescence

EnHD consists of one Tyrosine (Y), one Tryptophan (W) and 3 Phenyl Alanines (F). EnHD was excited at 280 nm. The fluorescence signal of EnHD upon thermal unfolding was affected by

(i) Contact quenching in the core of the native structure;

Quantum Yield (QY) of EnHD decreased with temperature at the measured low temperatures because of the quenching the fluorescence signal of Tyrosine and Tryptophan, by the nearby charged basic amino-acid residues like Lysines (K) or Arginines (R). This was followed by an increase in the QY of the protein with

temperature as the protein got unfolded and then showed a temperature dependence corresponding to a solvent exposed Tryptophan.

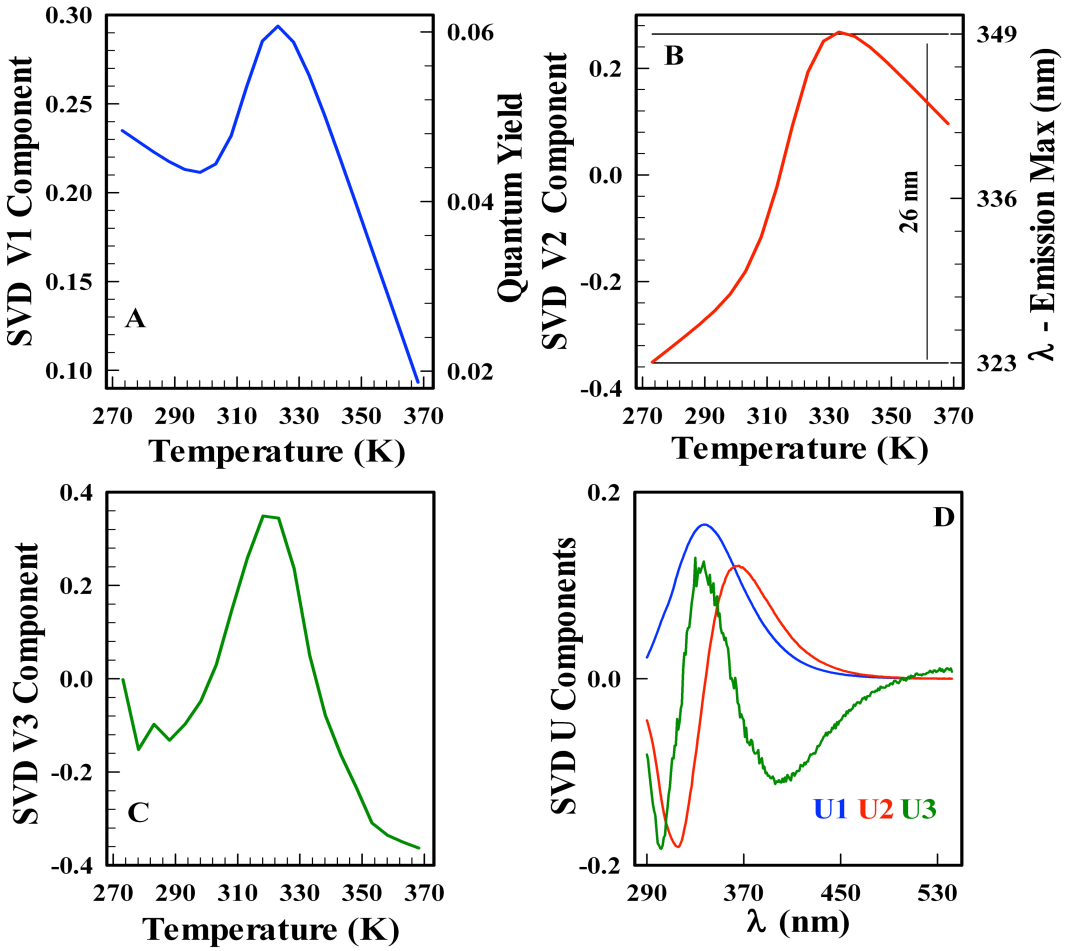


Figure 3-3. *SVD analysis of Fluorescence unfolding spectra. Figures A,B and C are the first three V components respectively from the analysis. The corresponding wavelength components U are shown in the figure D. In the figure A, V 1st component is compared with the Quantum Yield of the protein. In the figure B, V 2nd component is compared with the spectral shift.*

(ii) Solvent effects, resulting in large spectra shift of about 26 nm upon unfolding;

The wavelength emission maximum of the EnHD at native conditions was heavily blue-shifted (decrease in wavelength emission maximumum when compared to fully exposed W) implying less exposure of the aromatics towards the solvent at those conditions.

Tryptophan was buried in the aromatic core at native conditions whereas the Tyrosine residue was actually present in one of the loops of EnHD. This would probably imply the orientation of Try side-chain towards the aromatic core than towards solvent. At high temperatures, the emission spectrum of EnHD reached the emission spectrum of NATA implying a completely solvent-exposed Tryptophan.

(iii) Forster Resonance Energy Transfer from a Y as donor to a W^{87}, which are placed at a distance of ~12 Å in the native structure.

SVD Analysis of fluorescence emission spectra of EnHD with temperature revealed three significant components, with the second and the third components amount to ~ 20 % and ~ 1.2 % of the signal respectively with respect to the first component. The first two components compared very well with the quenching and the solvent effects respectively as described before. When the third wavelength component (U) coming from the SVD analysis was compared with the corresponding unfolding transition component (V): a) there was an increase in the signal transition (V) corresponding to the blue shift in the wavelength component (U); and, b) this was followed by a decrease in the signal transition (V) corresponding to the red-shifted peak in the wavelength component (U). This effect was intuited to be the FRET transfers between aromatic amino acid residues, as these residues are placed within their R_0 distance in the native structure. This effect could also be compared with an increase in the broadness of the fluorescence emission spectra with temperature followed by a decrease in the broadness, as the protein got unfolded.

3.3.4 Two-State Analysis of Thermal Unfolding Curves

Unfolding curves from multiple spectroscopic and calorimetry measurements were individually fit a two-state model. Unfolding curves at 222 nm, 280 nm and 1646 cm^{-1} from fCD, nCD and FTIR measurements were taken for the two-state analysis, whereas in the case of fluorescence, QY as a function of temperature was taken the analysis. Parameters obtained from the two-state analysis, enthalpy (ΔH_m) and melting temperature (T_m), were given in the table 3-1. Results from the two-state analysis showed a clear spread of melting temperatures (T_m) between different measurements. These values (T_m) compared very well with that of the melting temperatures obtained from the first-derivative of these unfolding curves. Differences in T_m from the two-state fit ranged from ~ 315 K for the average fluorescence signal to ~ 333 K to that of IR. Vant-Hoff Enthalpy (ΔH_m) obtained from such a fit revealed differences in the cooperativity. ΔH_m from two-state analysis implied a highly cooperative transition for fluorescence in comparison to that of other experiments.

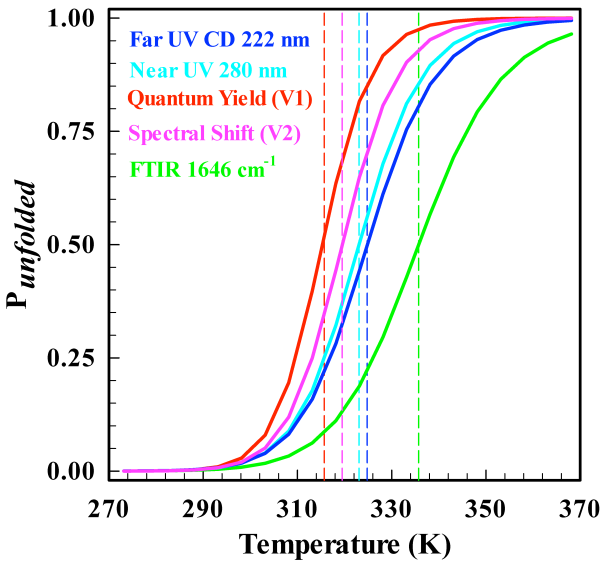


Figure 3-4. *Unfolding curves from different spectroscopic measurements represented as two-state probabilities.*

A global two-state analysis was performed on all the equilibrium thermal unfolding experiments for all the wavelengths (λ) in the case of spectroscopic measurements and for the heat capacity data in the case of DSC. While performing the global two-state analysis, ΔH_m and T_m were kept constant and all the folded and unfolded signals were allowed to vary for every λ. T_m of ~ 323 K and ΔH_m of ~ 129 kJ/mol were obtained from the fit. A statistical F-test was performed to identify whether the global or individual two-state fits best fit the experimental results. The inherent assumption of the F-test is that, a model with less number of parameters is in general statistically preferable. The calculcated value of the F ratio* is then used to estimate what is the probability that the current data has been produced by the simpler model (which in general produces a worse fit because it has fewer floating parameters). In this case, the F-test led to a p-value of zero, implying that the statistically simpler model (that is the global two-state model) is in fact inconsistent with the experimental data.

In order to see, if more sophisticated models were required to fit engrailed homeodomain data, both global three-state analysis and statistical mechanical model analysis of were performed on the data. In the global three-state analysis, ΔH_{m1} and T_{m1} and, ΔH_{m2} and T_{m2} were kept constant, and all the folded, unfolded signals and intermediate signals were allowed to vary for every melting curve. This analysis yielded the following thermodynamic parameters: ΔH_{m1} and T_{m1} of ~ 130 kJ/mol and ~ 319 K respectively and ΔH_{m2} and T_{m2} of ~ 118 kJ/mol and ~ 331 K respectively. This analysis was compared with the results from the statistical mechanical model and F-test again proved the statistical mechanical model (results from the best fit of the data to mean field

model shown later in this chapter) to be a better choice for the analysis of engrailed homeodomain results.

* $F-ratio = ((SLS1 - SLS2)/(N_{p2} - N_{p1}))/(SLS2/(N - N_{p2}))$ where 1 represents a model with less parameters and 2 represents a model with more parameters, N is the total number experimental data points, N_{p1} and N_{p2} are the number of fit parameters used in model 1 and 2 respectively and SLS1 and SLS2, the corresponding sum of least squares obtained from the fits.

Table 3-1. *Two State Fit Parameters*		
Probe	T_m (K)	ΔH_m kJ.mol^{-1}
Fluorescence QY (SVD V1)	315.3	161
Spectral Shift (SVD V2)	319.5	144
Near UV CD, 280 nm	323.1	130
Far UV CD, 222 nm	324.8	121
FT-Infrared, 1646 cm^{-1}	333.8	116

3.3.5 Estimation of Barrier Height to Folding from Calorimetry Data Using Statistical Mechanical Models

DSC unfolding curve was analyzed by two statistical mechanical models, Variable Barrier Model (VB Model) and One Dimensional Free Energy Surface Model (Mean Field (MF) Model) for various starting conditions/assumptions. A Bayesian analysis was used to rank the quality of the fit of the DSC data to both these models for all the different ways, the fits to these models were performed (Table 3-4, Table 3-5). In other words, a Bayesian analysis calculated a probability (P_h) for each fit result to assess the quality of the fit, according to the 'SLS' (sum of least squares) obtained from the fit to the model and how close the folded slope ($C_{p,N}$) was to the 'reference slope ($C_{p,Nr}$)'

chosen. The native slope obtained from the fit to DSC curve to two-state model was extremely steeper than the Freire's native slope.

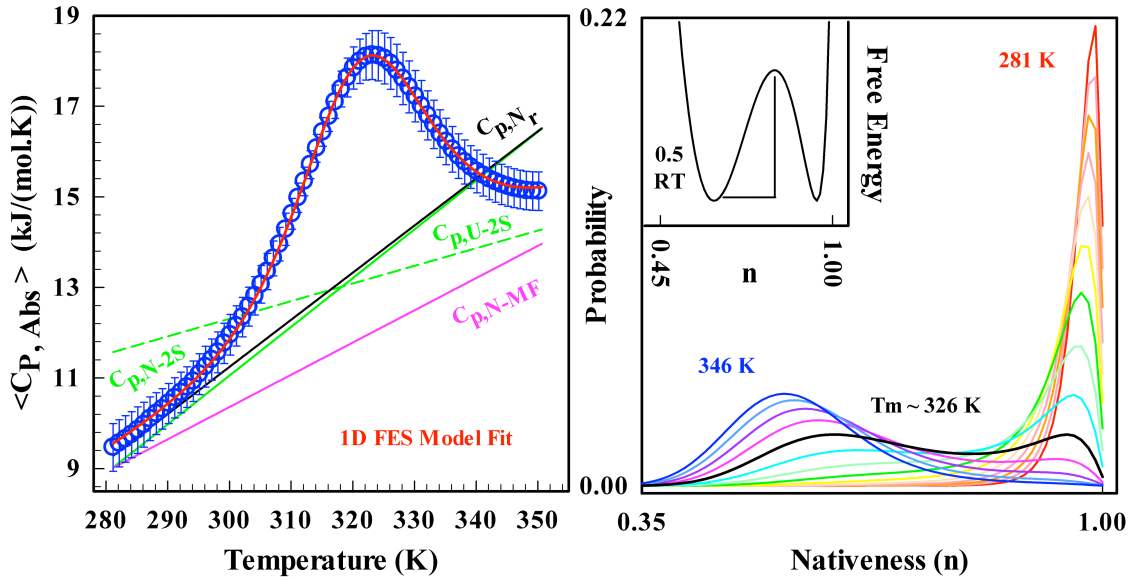


Figure 3-5. *On the left panel, the best fit (red line) of the DSC data (blue circles) to MF model, reference slope used in Bayesian analysis used to calculate as native baseline (black line) corresponding to molecular weight of engrailed, native baseline from the MF model fit (pink), folded (green line) and unfolded (green dashed line) from the two-state fit are shown. On the right panel, the probability distributions from the fit of the DSC data to MF model are shown. In the inset, free energy surface at the characteristic temperature is shown.*

Thus, when using Freire's native slope as reference slope, Bayesian analysis clearly ranked the fits that were worse, but with slopes close to Freire's native slope, higher in order. Thus, 'the average slope obtained from the two-state analysis of calorimetry data that were reported for a set of DNA-binding domains[88]' was chosen as the reference slope in this case (Table 3-3). Results/Parameters obtained from the fit of the data to both the models and from the Bayesian analysis were tabulated (Table 3-4, Table 3-5). The best fit of the MF Model to DSC data, based on the Bayesian analysis, yielded a barrier height to folding of ~ 1.28 kJ/mol (~0.5 RT) near the characteristic temperature (T_o) 326 K.

Table 3-2. *Parameters from the fit of DSC data to MF and Two-state models*					
Mean Field Model Fit Parameters				Two State Fit Parameters	
$\Delta C_{p,res}$ J.mol-1.K-1	$\Delta H_{loc,res}$ kJ.mol^{-1}	$\Delta H_{nonloc,res}$ kJ.mol^{-1}	$\beta(T_0)$ kJ.mol^{-1} (K)	T_m (K)	ΔH_m kJ.mol^{-1}
0	4.36	3.46	1.3 (326.2)	322.7	129

3.3.6 Global Fit of Multiple Probe Equilibrium Unfolding Measurements to the MF Model.

All the equilibrium thermal unfolding measurements were globally fit to MF model. A sigmoidal function was defined for the absolute value of the signal (<S>) vs. the order parameter, nativeness (n) for every λ.

$$< S(n,T) >= U + (F-U)\cdot\left(1 \Big/ 1+\exp(-c\cdot(n-n_m))\right) \tag{3.1}$$

In this equation, n is nativeness, F and U are folded and unfolded signals described as a function of nativeness, c is an apparent cooperativity parameter and n_m is the nativeness midpoint. This sigmoidal function used in conjunction with the MF model, were used to fit the entire unfolding data obtained from fCD, nCD, Steady-State Fluorescence and FTIR. In the case of fCD and nCD, no temperature dependence was assumed for the folded and unfolded signals vs. nativeness (n). In the case of FTIR, linear temperature dependence was assumed for the folded and unfolded baselines at every λ.

$$\begin{aligned} U(T) &= A + (T-273.15)\cdot B \\ F(T) &= C + (T-273.15)\cdot D \end{aligned} \tag{3.2}$$

In this equation, A and C are intercepts and B and D are slope parameters.

In the case of fluorescence, native and unfolded baselines were described as an exponentially decaying function as follows.

$$U(T) = E_u + F_u \cdot \exp(-(T - 273.15).G_u)$$
$$F(T) = E_f + F_f \cdot \exp(-(T - 273.15).G_f) \quad (3.3)$$

in which, E_u, F_u, G_u, E_f, F_f and G_f are phenomenological parameters used to model the data.

The maximum in the free energy surface between the two minimums is called transition state or the free energy barrier top. In the case of fitting the far UV CD and Near UV CD unfolding curve, the nativeness midpoint value was varied around transition state and between the nativeness values corresponding to two free energy surface minimums. The nativeness midpoint, yielding minimum SLS, was chosen at each wavelength. In the case of fluorescence and infrared, nativeness midpoint was fixed to 0.81, while fitting equilibrium spectra using MF Model as it corresponded to the nativeness value of the free energy barrier top. This was basically done to simplify the analysis of infrared and fluroscence that had more parameters in the fit to define the baselines.

3.4 Discussion

When the DSC unfolding curve was fit to a two-state model, folded and the unfolded baselines obtained from the analysis crossed near the melting temperature (> T_m) of the transition. As ΔC_p is the difference between the folded and the unfolded baselines, crossing of baselines would mean ΔC_p would increase and decrease depending

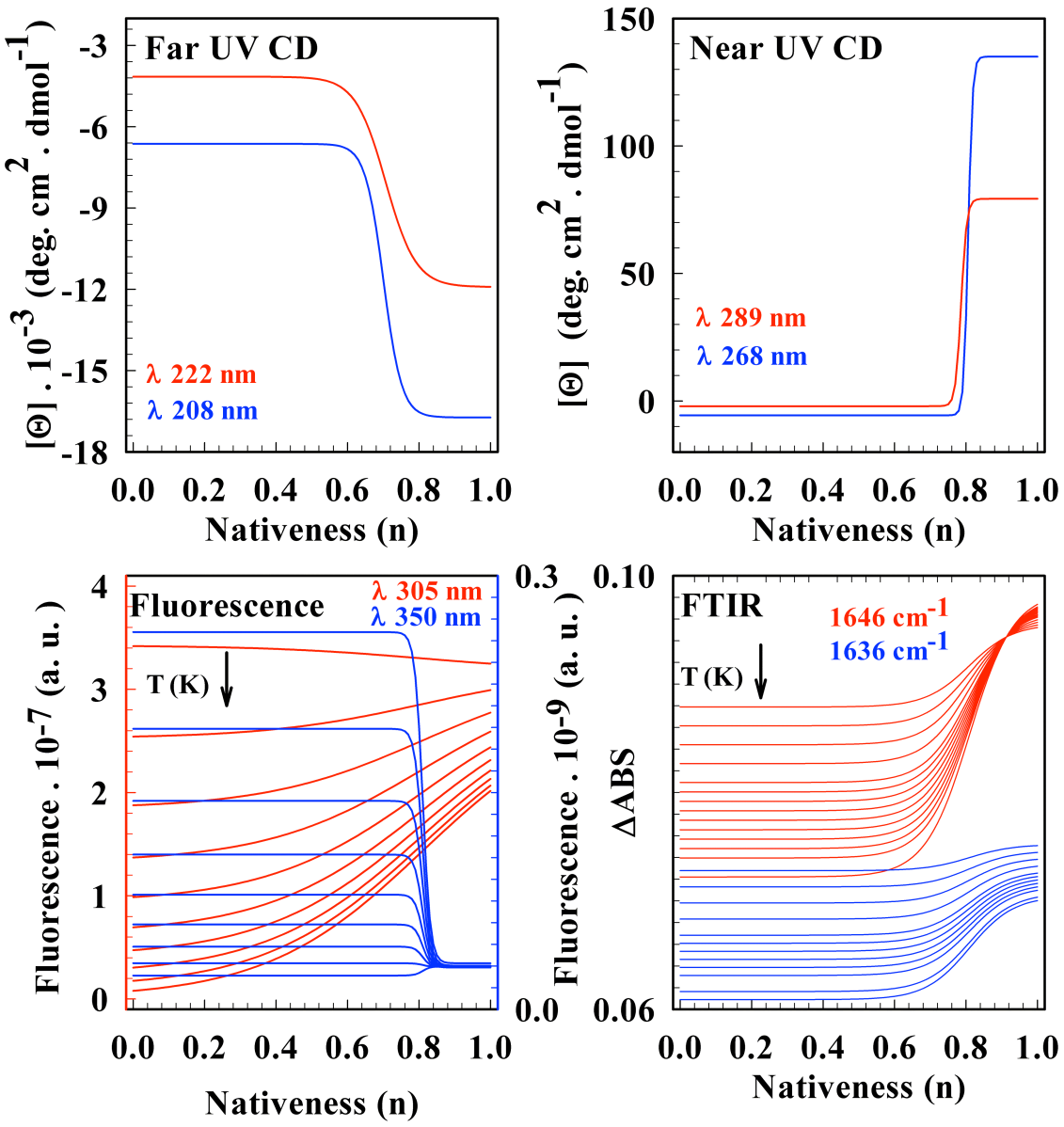


Figure 3-6. *Representative equilibrium signals as a function of nativeness for different spectroscopic probes. For far UV and near UV CD signals, no temperature dependence is taken into account, whereas for fluorescence and Infrared, temperature dependences are taken into account.*

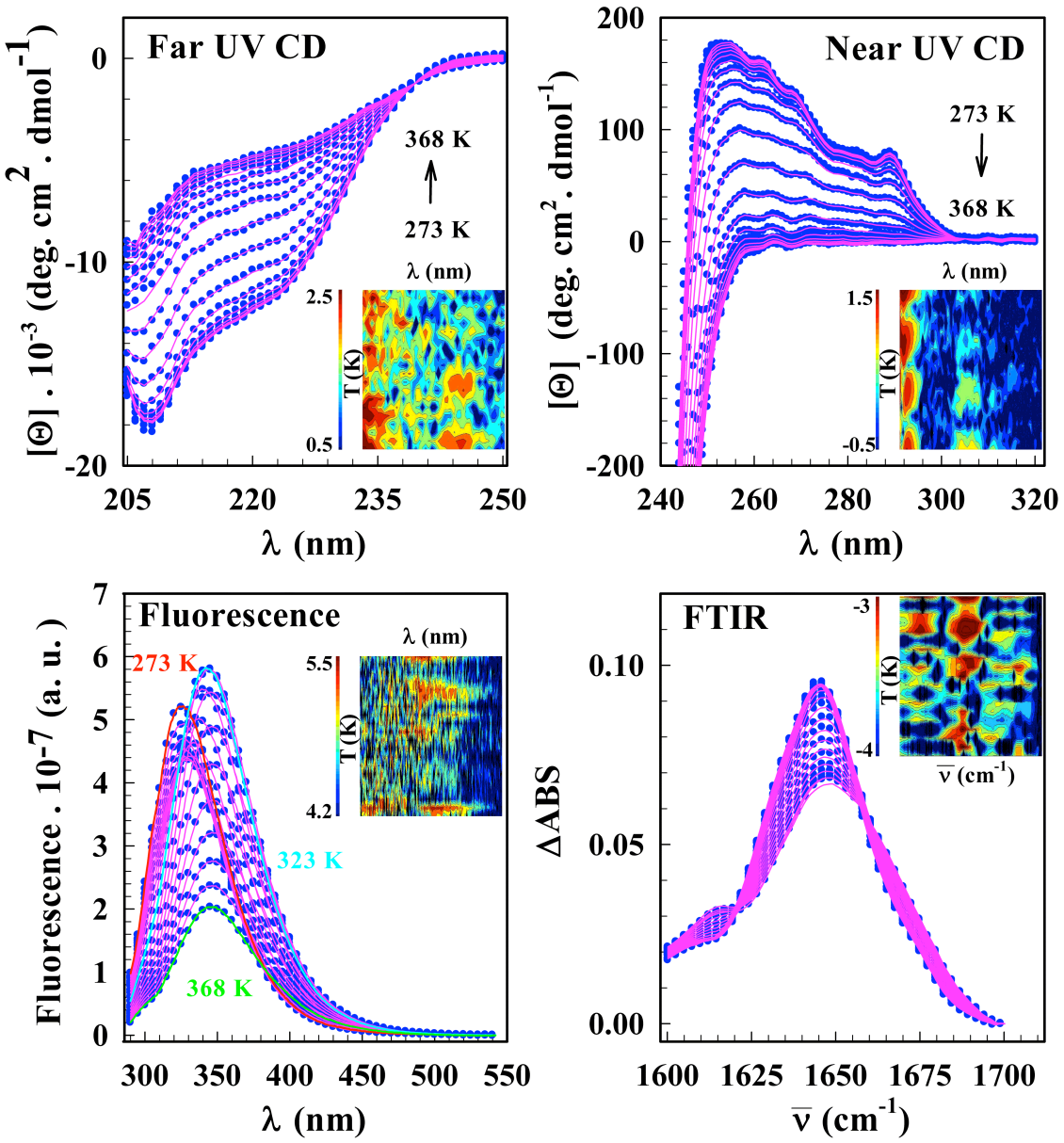


Figure 3-7. *Experimental unfolding spectra (blue circles) of Engrailed homeodomain from different spectroscopic probe. Fits to mean field model are shown in pink lines. In the inset of each graph, fit errors are shown in logarithmic scale. Color bars are shown along with the range of errors in each graph.*

on the temperature and make the estimation of this parameter complicated from this analysis. Broadness in the calorimetry transition and crossing of DSC folded and unfolded baselines from a two-state analysis clearly ruled out a two-state mechanism and have been shown to result in the presence of zero or marginal barrier height to folding[89]. Such crossing of baselines has also been observed in the DSC profiles of some DNA-binding domains, though have not been analyzed later for the estimation of barrier heights[88]. The baselines were not shown in the one of the earlier folding studies of engrailed by calorimetry and ΔC_p was reported to be 48 J/(K.mol.res)[53]. In another calorimetry study[60] performed on this domain, though the baselines were not shown, ΔC_p was reported to have calculated from the baselines from two-state analysis and was ~ 25 J/(K.mol.res) and this was two times lesser than previously reported value.

Two state analysis of Far UV CD measurement agreed with earlier studies. Though Near UV CD spectra[51] at native and unfolded conditions have been reported for this domain, full unfolding measurement has not been performed. Though fluorescence thermal unfolding measurement[51] was also performed, analysis of that transition has never been performed. In the current research work, thermal unfolding of engrailed was monitored by far UV CD, near UV CD, fluorescence and FTIR and all the unfolding curves were analyzed by a two-state model. Analysis of thermal unfolding curves by two-state model from these experiments clearly showed differences in the melting temperature and as well as in the cooperativity measure (enthalpy) between experiments. Thermal unfolding curves from just Infrared measurement showed wavelength-dependent unfolding transitions. In the case of fluorescence, a two-state analysis was performed just for the average fluorescence emission as well as for the first two components obtained

from the SVD analysis. While comparing the melting temperatures from two-state analysis of the first component with that of the second component, the average unfolding had a lower T_m in comparison to the unfolding curve obtained as a result of spectral shift. The third component from this SVD analysis that was intuited to be the FRET between W and Y implicated the decrease in the separation of these aromatic residues marked by the increase in the signal amplitude with temperature, followed by the increase in the separation of these residues marked by a decrease in the signal with temperature. It would be difficult to tell in this case which region of temperature marked the unfolding the protein when this result was compared with the results from the first two components. Global two-state analysis was just performed to demonstrate the parameters obtained from such an analysis would be an average representation of these distributed unfolding curves. Such differences in the melting temperatures have been shown as a characteristic of downhill folding in the cases of BBL and gpW[2,4,5]. This was demonstrated in the case of BBL both by monitoring the average behavior of different structural properties by low-resolution techniques and by monitoring atomistic unfolding of all amino acid residues by NMR.

In order to describe uniformly the unfolding behavior of engrailed accounting for all the complex unfolding observed, all the equilibrium measurements were then subjected to be analyzed by a statistical mechanical model. This could be done as the heat capacity data from calorimetry is directly related to protein partition function and also because such an analysis would be more realistic in this case as two-state probabilities could be represented solely for one particular unfolding curve, but the probability distributions produced by statistical mechanical model could account for all the unfolding

behaviors observed. Thus, DSC profile was analyzed by two statistical mechanical models for various criterions, and the best result for each model was identified. The best result from MF model was chosen as this model has been widely implemented for analyzing kinetic results and analyzing the rest of the data using the best result from the MF model would help extend the analysis for the kinetic results of this protein. Thus, the thermal unfolding data were sequentially globally fit to the MF model.

Equilibrium signals from different spectroscopic measurements were modeled as a function of the order parameter (nativeness, n), and this was implemented within the framework of Mean Field Model. Barrier heights to folding were calculated for every temperature and the barrier height near characteristic temperature (i.e., the melting temperature or the temperature at which in a bimodal distribution the peaks have same probabilities) was 0.47 RT. As per the classification of folding mechanisms based on the barrier height obtained near characteristic temperature, this comes under downhill folding mechanism. Another interesting result from this analysis was that it produced overlapping folding and unfolding baselines for the heat capacity and thus ΔC_p calculated was zero and this would imply an extremely broad unfolding transition. A higher value for enthalpy per residue from local contribution in comparison to non-local contribution support the view in which in alpha helices most stabilizing interactions come from the hydrogen bonding between two nearby peptide bonds and from the local side-chain interactions. As the temperature was increased, the folded populations decreased (peak near higher value of nativeness) and partially unfolded populations increased and this partially unfolded structures gradually unfolded to become fully unfolded conformations. This could be interpreted as, native conformations had to cross a short-barrier before

gradual unfolding. Native state heat capacity slopes obtained from both two-state analysis and Mean Field Model analysis were higher in magnitude in comparison to the Freire's native state heat capacity slope. Such higher values for slopes have been reported for DNA-binding proteins or domains when compared to globular proteins[60].

There were two attempts[3,68] to estimate the barrier height of engrailed from the calorimetry and kinetic data[53] reported previously for this protein. Both these studies reported low barrier, yet there were two reasons why the analysis on the fresh experimental data was considered. In the case of analyzing the previously reported calorimetry data using variable barrier model, as the calorimetry data reported were not reported in absolute heat capacity units, data had to be converted to absolute heat capacity units. This was done by translating the data based on Freire's baseline calculated for this protein according to the low temperature point and then a native baseline was derived from the low temperature point for this protein. For the kinetic analysis by mean field model, single exponential rate data were used and the analysis also inherently assumed a single exponential behavior while fitting the rate values.

If the folding mechanism of engrailed is two or three state with large barrier(s), it should have produced overlapping unfolding transitions. Thus, all these results, though cannot resolve multiple small barrier heights, can confirm the presence of a downhill folding behavior for engrailed homeodomain. Even with multiple barrier small barrier heights, the protein has to exert downhill behavior anyways.

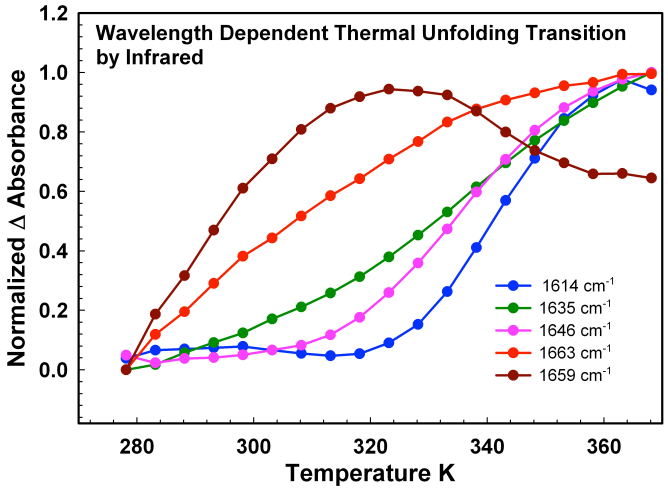


Figure 3-8. *Wavelength dependent unfolding behavior revealed by infrared.*

Table 3-3. *Slope values for different DNA binding domains*				
	DNA-Binding Protein/Domain	$b_{slope\text{-}protein}$ $J.K^{-2}.g^{-1}$	Molecular Mass $g.mol^{-1}$	$b_{slope\text{-}protein}$ / $b_{slope\text{-}Friere}$
1	Engrailed	0.0113	7583	1.6866
2	Mat α2	0.01244	9755	1.8567
3	Antennapedia	0.01287	8595	1.9209
4	LZ-GCN4	0.01522	8070	2.2716
5	HMGD-74	0.01339	8354	1.9985
6	HMG SOX5	0.01253	9085	1.8701
7	NHP6A	0.018	10708	2.6866
8	SRY	0.02304	10234	3.4388
9	Lef-79	0.02374	9310	3.5433
10	Zn-fingerTFIIIA	0.01504	12040	2.2448
			$\mu(b_{slope\text{-}protein})$	0.015757 $M_r.J.K^{-2}.g^{-1}$
			$\sigma(b_{slope\text{-}protein})$	0.004445 $M_r.J.K^{-2}.g^{-1}$
			Ratio = $\mu(b_{slope\text{-}protein})/(b_{slope\text{-}Friere})$	~ 2.35
			$b_{slope\text{-}Friere}$ = 0.0067 $M_r.J.K^{-2}.g^{-1}$	$\sigma(b_{slope\text{-}Friere})$ = 0.0013 $M_r.J.K^{-2}.g^{-1}$

Table 3-4. *DSC Analysis by VB Model*					
VB Model	**Baseline used**	**f**	**$\Sigma\alpha$ kJ.mol^{-1}**	**$a_{intercept}$ kJ.mol^{-1}.K^{-1}**	**b_{slope} kJ.mol^{-1}.K^{-1}**
1	Friere Baseline CpFol	0.24	7.7	8.74	0.0442
2	Friere Baseline CpFol + σ	0.11	185.7	9.10	0.0529
3	Friere Baseline CpFol - σ	0.3	265.1	8.38	0.0357
4	CpFol – Friere slope fixed	0.24	11.0	8.54	0.0442
5 (Rank 1)	Floating Baseline CpFol	1	141.0	7.63	0.0777
6 (Rank 4)	CpFol with gain (1.25) and offset (-2.9044)	0.57	51.9	8.02	0.0553

Table 3-4. *DSC Analysis by VB Model (Continued)*					
VB Model	**Baseline used**	**b/b_0**	**$\beta(T_0)$ kJ.mol^{-1}**	**SLS**	**P_h**
1	Friere Baseline CpFol	1	0 (318.5)	3.3406	0.0003
2	Friere Baseline CpFol + σ	1.194	4.11 (329.6)	1.4952	0.0138
3	Friere Baseline CpFol - σ	0.806	-1.51 (314.6)	2.2533	0.0011
4	CpFol – Friere slope fixed	1	0 (318.4)	1.1782	0.0133
5 (Rank 1)	Floating Baseline CpFol	1.756	0.47 (322.7)	0.0263	0.5129
6 (Rank 4)	CpFol with gain and offset	1.25	0 (319.9)	0.7774	0.0534

Table 3-5. *DSC Analysis by MF Model*						
MF Model	**Baseline Used**	**$\Delta C_{p,res}$ J.mol-1.K-1**	**$\Delta H_{loc,res}$ kJ.mol^{-1}**	**$\Delta H_{nonloc,res}$ kJ.mol^{-1}**	**$a_{intercept}$ kJ.mol^{-1}.K^{-1}**	**b_{slope} kJ.mol^{-1}.K^{-1}**
7 (Rank 2)	Floating Baseline	0	4.36	3.46	8.45	0.0712
8 (Rank 3)	Fixed $\Delta C_{p,res}$	10	3.97	3.66	8.69	0.0654
9	Fixed $\Delta C_{p,res}$	20	3.54	3.88	8.89	0.0597
10	Fixed $\Delta C_{p,res}$	30	3.04	4.15	9.06	0.0543
11	Fixed $\Delta C_{p,res}$	40	2.37	4.52	9.20	0.0491
12	Fixed $\Delta C_{p,res}$	50	1.19	5.25	9.35	0.0429
13	Fixed $\Delta C_{p,res}$	58	2.1211e-11	6.13	9.59	0.0380
14	Friere Baseline Fixed	53	1.3921e-12	6.10	8.74	0.0442
15	VB Floating Baseline Fixed	0	4.23	3.51	7.63	0.0777
16	VB Floating Baseline Stringent limits (as in 6)	52	5.2487e-08	6.09	8.02	0.0553
17	Friere Baseline - Fixed slope	43	2.07	4.70	9.37	0.0442
18	VB Baseline – slope fixed	0	4.37	3.45	8.18	0.0777
19	VB Baseline Condition (6) Slope fixed	23	3.40	3.95	9.06	0.0553

Table 3-5. *DSC Analysis by MF Model (Continued)*					
MF Model	**Baseline Used**	**b/b_0**	**$\beta(T_0)$ $kJ.mol^{-1}$ (K)**	**SLS**	**P_h**
7 (Rank 2)	Floating Baseline	1.608	1.28 (326.2)	0.2347	0.28673
8 (Rank 3)	Fixed $\Delta C_{p,res}$	1.476	2.32 (326.1)	0.8623	0.0781
9	Fixed $\Delta C_{p,res}$	1.348	3.87 (326.1)	1.8252	0.0117
10	Fixed $\Delta C_{p,res}$	1.227	6.35 (326.3)	3.0688	0.0011
11	Fixed $\Delta C_{p,res}$	1.109	10.905 (326.9)	4.5297	6.749e-05
12	Fixed $\Delta C_{p,res}$	0.969	23.0 (328.0)	5.8999	4.42e-06
13	Fixed $\Delta C_{p,res}$	0.858	41.1 (328.7)	20.761	4.64e-17
14	Friere Baseline Fixed	1	41.1 (328.2)	42.141	2.04e-32
15	VB Floating Baseline Fixed	1.756	1.58 (325.8)	37.71	1.91e-28
16	VB Floating Baseline Stringent limits (as in 6)	1.25	41.0 (327.6)	93.625	1.40e-69
17	Friere Baseline - Fixed slope	1	13.5 (327.5)	5.3872	1.152e-05
18	VB Baseline – slope fixed	1.756	1.23 (325.8)	1.8485	0.0242
19	VB Baseline Condition (6) Slope fixed	1.25	4.50 (326.3)	2.4658	0.0031

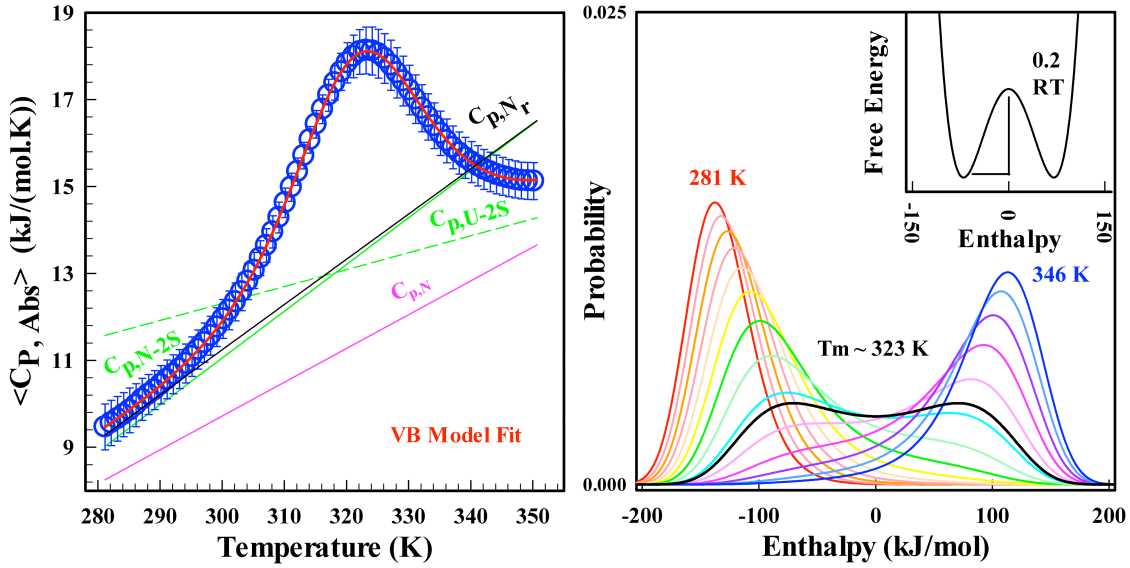


Figure 3-9. *On the left panel, the best fit (red line) of the DSC data (blue circles) to VB model, reference slope used in Bayesian analysis used to calculate as native baseline (black line) corresponding to molecular weight of engrailed, native baseline from the VB model fit (pink), folded (green line) and unfolded (green dashed line) from the two-state fit are shown. On the right panel, the probability distributions from the fit of the DSC data to VB model are shown. In the inset, free energy surface at the characteristic temperature is shown.*

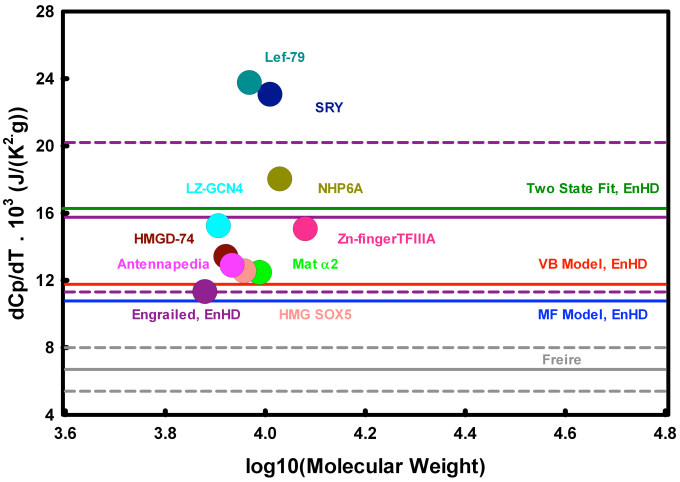


Figure 3-10. *Native state heat capacity slope values for different DNA-binding domains[88]. Freire's native state slope value, slope values from two state fit, best VB Model fit and best MF model fits are indicated.*

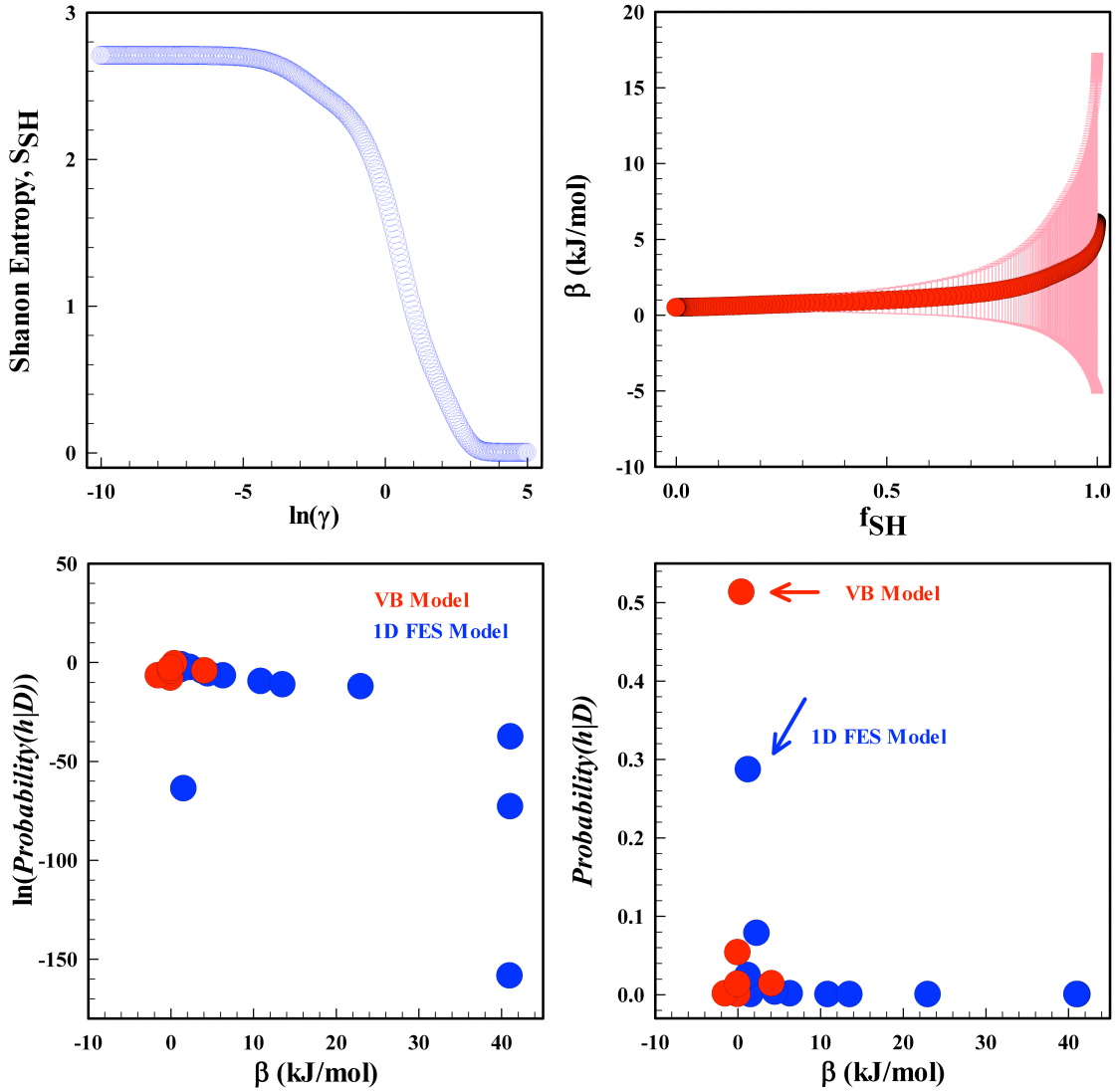


Figure 3-11. *Bayesian analysis results from the fit of DSC data to VB and MF models for various criterions. Top left panel graph shows the plot of Shannon entropy vs. ln(ν). Top right panel shows the barrier height for different values of relative Shannon entropy with the error bars. Bottom panels show the Bayesian probability for the barrier heights estimated from fits of DSC data to these models for several criteria (see table 3-3, 3-4). The best results for both VB and MF models are indicated.*

Chapter 4: (Un)folding of Engrailed Homeodomain studied by double perturbation measurements

4.1 Abstract

Studying the (un)folding of single domain proteins by both denaturant and temperature using a single spectroscopic probe has been used to distinguish downhill and two-state mechanisms. Here, we study an ultrafast folding engrailed homeodomain that has been reported to exert a three-state folding mechanism, by a double perturbation measurement using far UV CD. Results from the analysis of double perturbation measurement reveal complex coupling between the denaturant and temperature. This result cannot by itself say that this is the signature of downhill behaviour in this particular case of engrailed homeodomain. Non-coincidental unfolding of engrailed by additional chemical unfolding measurements by multiple spectroscopic probes at a series of temperature marks a downhill behaviour and also confirm that complex coupling has resulted out of that behaviour.

4.2 Introduction

Calculation of heat capacity has traditionally been done by analyzing the thermal melting curve from differential scanning calorimetry. Double perturbation experiment (unfolding monitored by both temperature and urea, for example) has also been used to evaluate the heat capacity of proteins[90]. Although heat capacity can be obtained by a analyzing a sigmoidal thermal denaturation curve by a two-state model, heat capacity value obtained from the global two-state analysis of double perturbation experiment is considered more reliable. This is because (folded and) unfolded signal can unambiguously be represented and could well represent the entire folding in comparison to a single curve for the estimation of heat capacity.

In the case of identifying complex folding scenarios, typically more spectroscopic probes are used. But, double perturbation measurement (temperature vs. urea) was reported to distinguish between two-state and downhill folding scenarios using a single spectroscopic probe[7]. This was demonstrated[6,4] in the cases of BBL and gpW by using far UV CD as a spectroscopic probe using temperature and urea as means to perturb the system. The rationale behind this test was that, in the case of two-state scenario, the positions (reaction coordinate) of folded and unfolded wells in a free energy surface would not change with urea, where as the positions of these wells would change with urea for a downhill folding scenario. In the case of globally downhill folding protein, there would be one minimum that would change with urea concentration. Thus, a non-linear relationship was observed between enthalpy and urea concentration in the case of downhill folding mechanism as a contrary to linear correlation between enthalpy and urea that was assumed and observed in cases of two-state proteins.

Engrailed homeodomain, as it was reported to exert a three-state mechanism in earlier studies, both double perturbation measurement and multi probe chemical unfolding measurements were performed. If this protein has two large barriers like how it has been interpreted, the positions of folded, unfolded and intermediate well wouldn't move with urea concentration. If this protein has two small barriers or a single barrier, the positions of these wells would move with urea. In both these scenarios, it could be difficult to interpret the results of double perturbation measurement using a single probe for a protein showing sigmoidal transitions. But, if it is used in combination with multiple probe chemical unfolding measurements, the results from these measurement can explain the double perturbation experiment results from some perspectives.

4.3 Results

4.3.1 Thermal Unfolding – Double Perturbation Experiment by Far UV CD

A series of thermal unfolding curves from far UV CD measurements were obtained in the temperature range between 273 K and 368 and the urea concentration between 0M and 5M. There was no cold denaturation observed in these experiments. A global two-state fit was performed on this double perturbation experimental data. For the global analysis,

a) a linear temperature dependence was assumed for the folded baseline,

$$N = S_{N0} + S_N .(T - T_{ref}) \tag{4.1}$$

where S_{N0} and S_N were folded intercept and the slope that described the dependence of folded signal on temperature,

b) a linear dependence on urea and a term describing the coupling of urea binding with respect to temperature was assumed in the case of unfolded baseline,

$$U = S_{U0} + S_{U1}.[D] + ((S_{U2} + S_{U3}.[D]).(T - T_{ref}) \quad (4.2)$$

where S_{U0} and S_{U1} were unfolded signal intercept and its dependence on denaturant, S_{U2} and S_{U3} are parameters that described the urea binding to protein coupled to temperature dependence, and [D] was the denaturant concentration, and

c) the thermodynamic parameter, ΔH_m, was assumed to vary linearly with temperature as it should be the case for a perfect two-state folding scenario.

These descriptions of baselines and thermodynamic parameter fit the data reasonably well. A closer introspection of the fit revealed the fits were good at lower concentrations of urea and reasonable at higher concentrations of urea and slight discrepancies occurred at the lowest temperature measured for the experiments in the mid-range urea concentraions. Global analysis estimated three thermodynamic parameters, T_m, ΔH_m and ΔC_p. An estimation of $\Delta C_{p,res}$ from these fits was 34 $J.mol^{-1}.K^{-1}.res^{-1}$. Melting temperature and ΔH_m obtained from this analysis were 325 K and 140 $kJ.mol^{-1}$ respectively. This was followed by fitting the thermal unfolding curves individually to a two-state model. ΔC_p values varied from 1.65 $kJ.mol^{-1}.K^{-1}$ to 0.2 $kJ.mol^{-1}.K^{-1}$ from the lowest to the highest concentration of urea concentration used for the measurement. ΔH_m calculated from individually analyzing the chemical unfolding curves between 273 K and 368 K by a two-state analysis was clearly non-linear. Global fit overestimated the enthalpy at 298K (ΔH_{298K}) at the concentrations of urea below 1 M, underestimated it at 1M, was linear at

the ranges between 1.5 M and 2.5 M and heavily overestimated at concentrations greater than 3 M.

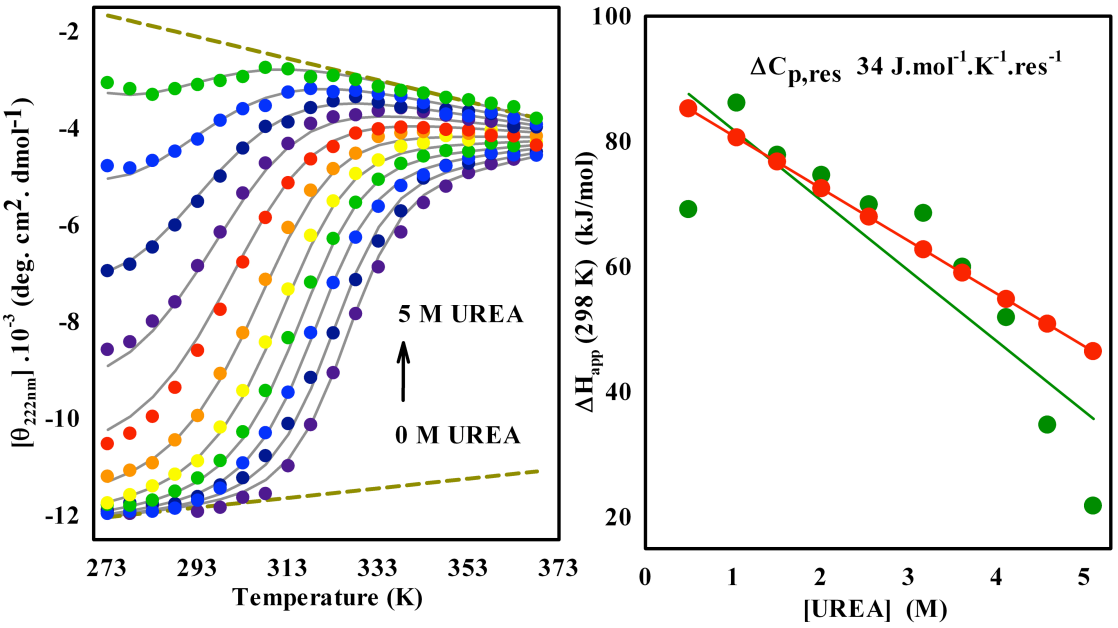
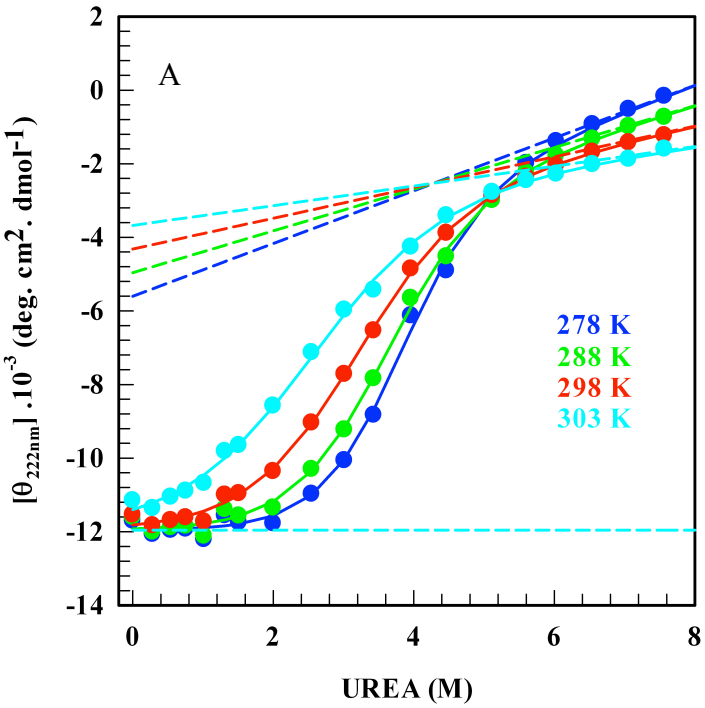
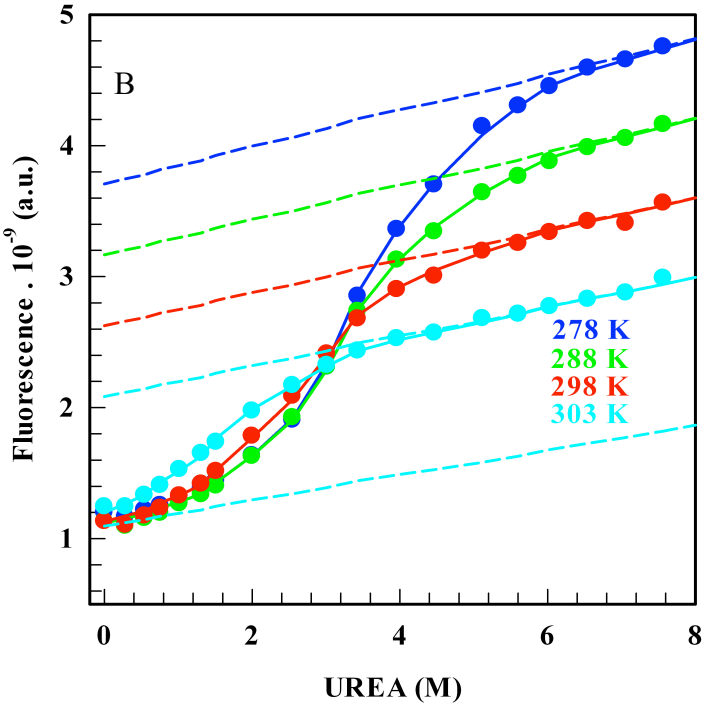


Figure 4-1. *On the left panel, global two-state fit (grey) of the experimental data (circles) are shown. On the right panel, apparent enthalpies at 298K from the global two-state fit are shown in red, apparent enthalpies at 298 K estimated from individual two-state fit are shown in green circles and the corresponding linear fit is shown as a green line.*

4.3.2 Chemical Unfolding by Steady-State Fluorescence, Far UV CD and Bulk FRET

A series of chemical unfolding experiments on the unlabeled protein by far UV CD and steady state fluorescence and a series of chemical unfolding measurements on the engrailed homeodomain doubly labeled with Alexa 488 and Alexa 594 to measure the bulk FRET of the protein were done in the temperature ranges between 278K and 308 K. Urea concentrations between 0M and 8M were used for the far UV and Steady-state fluorescence measurements on unlabeled protein. To measure bulk FRET on fluorescent-labeled protein, urea concentrations between 0M and 7M were used. Results obtained


A
278 K
288 K
298 K
303 K
[θ222nm] .10-3 (deg. cm2. dmol-1)
UREA (M)


B
278 K
288 K
298 K
303 K
Fluorescence . 10-9 (a.u.)
UREA (M)

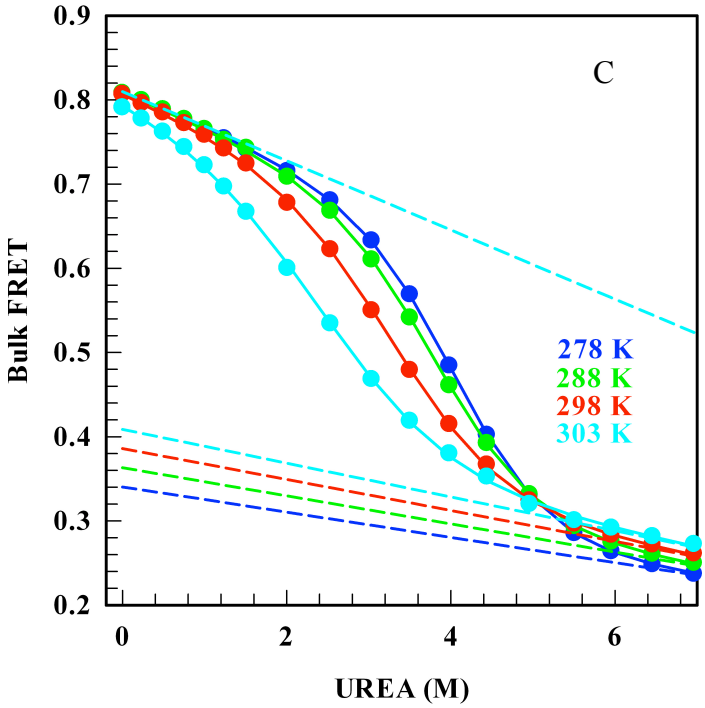


Figure 4-2. *Chemical unfolding measurements (data in circles) at a series of temperature by far UV CD (A), fluorescence(B) on unlabeled EnHD and FRET measurements on the EnHD labeled with Alexa 488 and Alexa 594 at the termini.*

from each of these spectroscopic probes were globally fit for the entire temperature range by a two-state model. In this global analysis, stringent conditions were imposed on the baselines while analyzing the curves from all the four temperatures together, while allowing the thermodynamic parameters to vary.

For the global two-state analyses, baselines were described as follows.

i) linear dependence on urea for the folded baseline

$$N = S_{N0} + S_N \cdot [D] \tag{4.3}$$

where S_{N0} and S_N were folded intercept and the slope that described the dependence of folded signal on denaturant and [D] was denaturant concentration, and

ii) dependence of binding of urea to protein with temperature for the unfolded baseline.

$$U = S_{U0} + S_{U1}.(T - T_{ref}) + ((S_{U2} + S_{U3} \cdot (T - T_{ref})) \cdot [D]) \quad (4.4)$$

where S_{U0} and S_{U1} were unfolded signal intercept and its dependence on temperature, S_{U2} and S_{U3} are parameters that described the urea binding to protein coupled to temperature dependence and [D] was denaturant concentration. Such analyses resulted in good fit of the data. Chemical denaturation midpoint was a) 3.7 M at 278 K and 2.5 M at 308 K for Far UV CD, b) 3.5 M at 278 K and 1.5M at 308 K for steady state fluorescence and c) 4.1 M at 278 K and 2.6 M at 308 K for bulk FRET from such a global two-state fit analysis. m-values and ΔG_{H20} from these analyses were reported in the table 4-1.

Table 4-1. *Two State Fit Parameters – Chemical Unfolding Measurments*

	Far UV CD			Fluorescence			Bulk FRET		
T(K)	C_m (M)	m-value $kJ.mol^{-1}.M^{-1}$	ΔG_{H20} $kJ.mol^{-1}$	C_m (M)	m-value $kJ.mol^{-1}.M^{-1}$	ΔG_{H20} $kJ.mol^{-1}$	C_m (M)	m-value $kJ.mol^{-1}.M^{-1}$	ΔG_{H20} $kJ.mol^{-1}$
278	3.7	3.90	15	3.5	3.13	11	4.1	3.97	16
288	3.5	3.54	12	3.3	3.32	11	3.9	3.82	15
298	3.1	3.19	10	2.7	3.50	9	3.4	3.52	12
308	2.5	2.75	7	1.5	3.58	5	2.6	3.12	8

Results from fluorescence experiments on the unlabeled protein revealed more complexities. The average fluorescence signal of the protein was affected by i) solvent effects leading to the spectral shift of ~ 24 nm between extremes of temperature and urea denaturing conditions and ii) Forster Resonance Energy Transfer between Tryptophan and Tyrosine that are placed within the R_0 of ~ 14 Å. Basically, singular value decomposition analysis of the entire data (all 4 temperatures) revealed three components, the first corresponding to average fluorescence signal, the second corresponding to the spectral shift and the third one intuited to be the FRET between aromatic amino acid

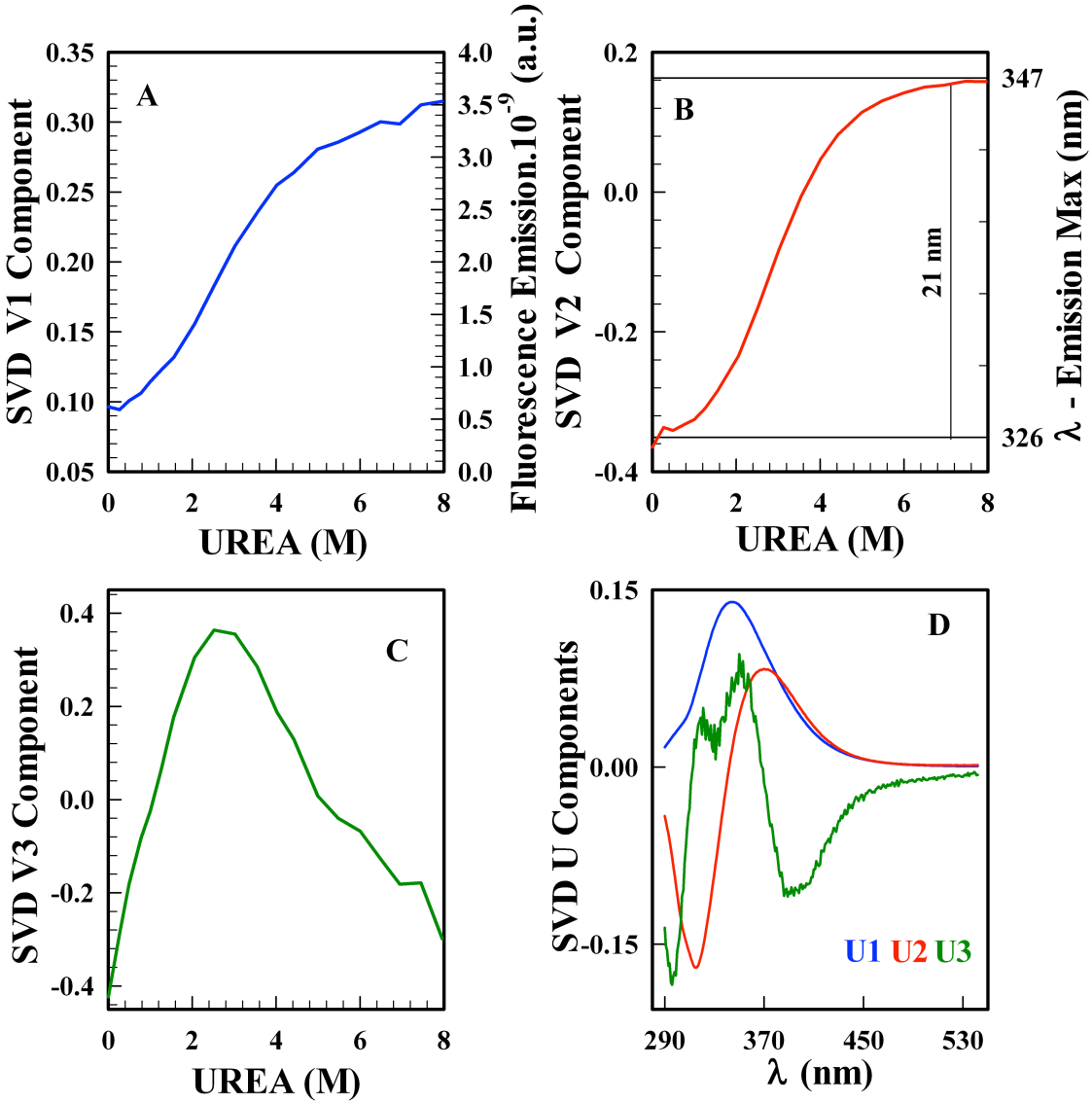


Figure 4-3. *Global SVD analysis of Fluorescence unfolding spectra obtained at all the temperatures and all the concentration of denaturants. Figures A,B and C are the first three V components respectively from the analysis. The corresponding wavelength components U are shown in the figure D. In the figure A, V 1st component is compared with the Fluorescence emission of the protein. In the figure B, V 2nd component is compared with the spectral shift.*

residues. The last two components contributed 13 % and 0.56 % with respect to the first component. These effects, when the data were analyzed at each temperature, contributed to ~ 15 % and 0.53 % respectively at 278 K and decreased to about 9% and 0.27 % respectively at 308 K. Difference in the spectral shift, i.e. differences in the emission

maximum between extremes of chemical unfolding conditions at a particular temperature, was about 23 nm at 278 K and this decreased to ~ 15 nm at 308 K.

4.4 Discussion

Results from the chemical unfolding measurements on EnHD revealed the following:

a) Differences in the apparent chemical denaturation midpoints (Cm) between different spectroscopic probes. Estimation of chemical denaturation midpoints by global two state analyses were around 3.1 M by Far UV CD and 2.7 M by Steady State Fluorescence on the unlabeled protein sample, and 3.4 M by Bulk FRET on the fluorescent-labeled protein sample at 298 K. Such differences were observed in other temperatures as well.

b) Decrease in the cooperativity in the chemical unfolding from 5^0C to 35^0C. Cooperativity in the unfolding transitions can be evaluated in terms of broadness in the unfolding transitions or from how sigmoidal the unfolding curves are. In all the measurements, the chemical unfolding curves at 308 K were comparatively broader than the unfolding curves obtained at 278K. Despite the fact cooperativity in the unfolding transition decreased with temperature, the unfolding transitions were still broad even at the lowest temperature measured. To simply illustrate this even without considering how much sigmoidal the transitions were, bulk FRET of the labeled protein sample spanned broad range of FRET values between 0.81 and 0.24 at 278 K to 0.79 and 0.27 at 308 K.

c) Anomalies in the trend of ΔG_{H20} and m-values. C_m obtained from two-state analysis decreased with increase in temperature as expected. m-value is the energy required to expose the protein to the solvent when the concentration of denaturant got increased by

1M. In other words, this m-value is directly related to solvent accessible surface area. ΔG_{H20} is the stability parameter. In the case of Far UV CD measurements on the unlabeled sample and bulk FRET on the labeled sample, both the m-values and ΔG_{H20} values decreased with increase in temperature. But, in the case of steady state fluorescence: i) ΔG_{H20} values remained the same at 278 and 288 K and decreased from 288 K to 308 K. ii) m-values increased with increase in temperature. Fluorescence signal first decreased and then increased with temperature. Decrease in the average fluorescence signal was because of the quenching of fluorescence signal by the charged amino acid residues near aromatic residues. This affected the estimation of thermodynamic parameters, m-value and ΔG_{H20}. Difference in the thermodynamic parameters between fluorescence and CD chemical unfolding measurements at room temperature were also observed in a study before for this domain[50]. This was interpreted to be a multi-state mechanism. But, the nature of multi-state was not explained, in the sense whether barriers are assumed larger or marginal. However, the differences in the C_m between three probes at four temperatures clearly should be a mark of a downhill behavior or the existence of low-barriers.

Analysis of chemical unfolding measurements on unlabeled engrailed revealed non-overlapping unfolding transitions when the unfolding represented by average fluorescence signal was compared with the unfolding represented by wavelength emission maximum. SVD analysis also revealed a third component intuited to be FRET between W and Y. It would make sense that FRET decreased with temperature. For the FRET, the initial increase in signal amplitude with urea implied the decrease in the distance between W and Y followed by a decrease in the amplitude would imply W and

Y started to move apart. Such complex behaviors could be observed merely from a single spectroscopic probe for engrailed. This also implies a downhill behavior.

Global analysis of double perturbation measurement by far UV CD led to an estimation of $\Delta C_{p,res}$ and the value obtained of 34 $J.mol^{-1}.K^{-1}.res^{-1}$ was less than 58 $J.mol^{-1}.K^{-1}.res^{-1}$ reported from Robertson and Murphy data set[91]. ΔH_m calculated from individually analyzing the chemical unfolding curves between 273 K and 368 K by a two-state analysis was clearly non-linear and had a characteristic curvature. This implied a complex coupling between temperature and urea for the unfolding behavior of engrailed. Such complex coupling was taken as a signature to imply downhill folding behavior in the cases of BBL and gpW and made sense in those cases. This rules out two-state mechanism anyways. But, as folding of engrailed was described by a three-state mechanism in earlier studies, this criterion can no longer be a decisive criterion for the case of engrailed homeodomain. This result has to be complemented by other experimental results. If the folding of engrailed proceeded with a three-state mechanism or multi-state mechanism with larger barriers, it should any ways produce sharp overlapping unfolding curves between different spectroscopic probes. If this was what observed, then the complex coupling in the double perturbation measurement resulted due to multi-state mechanisms with significant barriers. In our case, broad non-overlapping unfolding behaviors observed in chemical unfolding at four low temperatures clearly suggest that the characteristic curvature observed in double perturbation measurement actually resulted due to the downhill nature of unfolding for engrailed homeodomain.

Chapter 5: Multiple spectroscopic probes monitored extremely complex fast folding kinetics of Engrailed Homeodomain

5.1 Abstract

Earlier kinetic studies on engrailed homeodomain using a single spectroscopic probe showed the presence of a faster phase in addition to the slow folding phase and this was analyzed by a conventional three-state model. But, multiple-probe thermal unfolding measurements of engrailed homeodomain, showed the signatures of a downhill folding behavior, and as well confirmed such a behavior by estimating the barrier height.

Here, we study the kinetics of EnHD using two spectroscopic probes, nanosecond resolution Infrared and Fluorescence Laser Temperature Jump Kinetic Measurements. Decays from Infrared kinetics monitored at 1636 cm^{-1} & 1646 cm^{-1} are clearly exponential, where as spectral decays from Fluorescence are non-exponential kinetics and these non-exponential decays are fit to double exponential. Furthermore, SVD analysis of Fluorescence decays shows three significant components corresponding to average fluorescence signal, spectral shift and intuited FRET between W and Y. The slow folding phase from both the experiments agree with each other and these relaxation rates were in tens of microseconds. Kinetic amplitudes for the slow folding phase show probe-dependency. This implies a downhill folding mechanism. As engrailed folds with a low barrier height, additional faster phase observed is due the transitions happening in the transition state region and is called 'molecular phase'.

Here, we analyze the kinetics by extending the results from equilibrium and then globally fit all the experimental decays to statistical mechanical model. We calculate

diffusion coefficients at every measured temperature.

5.2 Introduction

Proper folding of proteins is important for proper functioning of proteins. For a long time and even until now, many protein folding studies have been performed with the notion that folding between an extended chain and the native state is equivalent to a chemical two-state kinetics and the chance that any transient state can be observed in between becomes negligible, as it is assumed to have a very high energy with respect the ground state and hence little probability to observe such a state. The difference in energy level between native state and the transition state is called the 'barrier height to folding'. If this barrier height is actually low, intermediate states can be observed. Actually, proteins with a low barrier height to folding would continuously unfold and therefore, the entire folding path can be monitored.

The problem in analyzing the experimental data especially for a low-barrier scenario with conventional models (in this case two-state) can be illustrated. Both the proteins that fold with a small barrier and the ones that fold with a larger barrier would produce sigmoidal unfolding transitions and could be fit by a two-state model. The rate data can also be easily fit using conventional models. In terms of what the average signals and rate data represent for both these scenarios, it could be explained as follows. In the case of large barrier, the signal value is the average value of the signals contributed by folded and unfolded populations. In the case of low barrier or no-barrier, for thermal denaturation measurements, the signal value is the average value from folded conformations at low temperature, average value from the intermediate conformational

ensemble at any intermediate temperatures and average value from the unfolded conformations at high temperature. In the case of kinetics, the rate of exchange is between folded and unfolded conformational ensembles for a large barrier, and it is described between two intermediate conformational ensembles in the case of no barrier. In the case of low barrier, even the native state and unfolded ensembles move upon denaturation. Thus, it can be interpreted similar to the case of no barrier. Thus, for both equilibrium and kinetics, it is clear that conventional models do not capture the underlying mechanism for the low barrier (downhill) scenario. Statistical mechanical models prove to be more useful for analyzing the results in such a scenario.

Downhill folding scenarios have been characterized and identified by same rate of exchange but non-coincidental unfolding between different spectroscopic probes[23]. Though non-exponential (stretched relaxation) relaxation has been observed in downhill folding[92-94], it is not a strict signature for the same[95]. Many equilibrium, kinetic[49-57] and MD simulations have been performed on engrailed homeodomain[75-84]. Multiple probe thermal unfolding of engrailed reveals the presence of downhill folding mechanism in this protein. Though some MD simulations support three-state nature of folding of EnHD, there's another study that states the inability[84] to perform MD simulations of EnHD using the same crystal structure. Kinetic studies by fluorescence temperature jump reveal EnHD is an ultrafast folding protein. The presence of additional faster phase was interpreted with an on-pathway model. But, theoretical studies predict small barriers for this protein. In order to demonstrate which folding mechanism EnHD resort to by kinetic experiments, multiple probe temperature jump studies will need to be performed. Results

obtained from these measurements and reasons for the possible origin of the fast phase will be discussed.

5.3 Results

5.3.1 Thermal Unfolding - Infrared T-Jump

Equilibrium Infrared Measurements on EnHD revealed two peaks corresponding to α-helical signal in the Amide I region, one at 1646 cm^{-1} and another at 1636 cm^{-1}. These two frequencies monitored the buried and exposed α-helical signals. Equilibrium data clearly revealed a more co-operative unfolding for 1646 cm^{-1} than for 1636 cm^{-1}. Unfolding at 1636 cm^{-1} preceded unfolding at 1646 cm^{-1}. Kinetics of EnHD was monitored by IR T-Jump measurements at these two frequencies, spanning the temperature range between 303 K and 343 K, from 800 ns.

Maximum change in the signal amplitude at 1636 cm^{-1} was followed by the maximum change in the signal amplitude at 1646 cm^{-1} conforming to the equilibrium results. The difference in T_m between these two frequencies was not significant, yet in terms of the order of the process, it would make sense, a less exposed α-helical signal should precede buried α-helical signal. Maximum signal change occurred around 334 K. Amplitudes at both the wave numbers were broader. IR decays at both the frequencies at all the temperatures were clearly exponential decays. Observed relaxation rate estimated from the fit to single exponential was 10 μs at the lowest T measured to 1.1 μs at the highest T measured.

5.3.2 Thermal Unfolding - Fluorescence T-Jump

SVD analysis of thermal unfolding data obtained by monitoring the fluorescence of EnHD revealed three components. These three components correspond to average fluorescence signal, spectral shift and FRET between aromatic amino acid residues respectively. EnHD was excited at 288 nm and ~ 6.8 K T-Jump was produced on the protein sample. Spectral decays were acquired between 285 K and 337 K, from 10^{-8} s to 10^{-3} s. Global SVD analysis of all the relaxation decays obtained between 285 K and 337 K revealed the same three components as was observed in equilibrium measurements. These three components simultaneously monitored the unfolding kinetics of EnHD. The spectral shift and FRET components amounted to about 14 % and 1.2 % signal with respect to the first component or the average signal. These conformed reasonably well with the equilibrium results obtained, in which the solvent and FRET effects were 17% and 0.8 % with respect to the average signal between the same temperature ranges.

From the observed decays, it was clear that most of decays were non-exponential. Though some decays looked more stretched than a double exponential, it looked more appropriate to use a double exponential to uniformly analyze the data than a stretched exponential. This was because of the presence of few decays in which the signals went in both positive and negative directions. Decays were analyzed any ways by also stretched exponential in order to gain insight about the extent of non-exponential behaviour with temperature.

Relaxation obtained at each final temperature was globally fit for the three components at each temperature. In other words, for a double exponential fit, two rates and 3 amplitudes for the three components per kinetic phase were used to fit the three

decays of the three components at one final temperature simultaneously. The fast rate was fixed to 0.8 μs and this was basically done to keep the fast phase near to the folding speed limit while analyzing the data. Relaxation rates for the slow major folding phase were between 25 μs at 285 K and 1.3 μs at 337 K. Maximum signal change for the slower folding phase for the second component or for the spectral shift occurred at ~ 326 K, whereas the maximum amplitude change corresponding to the average fluorescence signal or the first component happened at ~ 323 K. It was difficult to interpret the signal change for the third component. In the case of amplitudes from the fast phase, they contributed very less signal in comparison to the slow phase and also depending upon the component.

When the decays were fit to a stretched exponential* globally for the three components at each temperature, the fit yielded a β value of ~ 0.46 at 285 K implying a highly stretched relaxation at that temperature, and the β value obtained from the fit kept increasing with temperature and reached a β value of 1 at 337 K implying a single exponential relaxation at that temperature.

*Stretched exponential equation is given by $y = a_0 + A \cdot \exp\left(-(t/\tau)^{\beta}\right)$, where t is the time, τ is the relaxation rate and β is the stretched exponential exponent.

5.3.3 Analysis of Decays from IR and Fluorescence T-Jump Kinetics Using One Dimensional Free Energy Surface Model

As the results from multiple probe kinetic experiments were complicated, in which one probe showed non-exponential behavior whereas the other did not, the analysis were

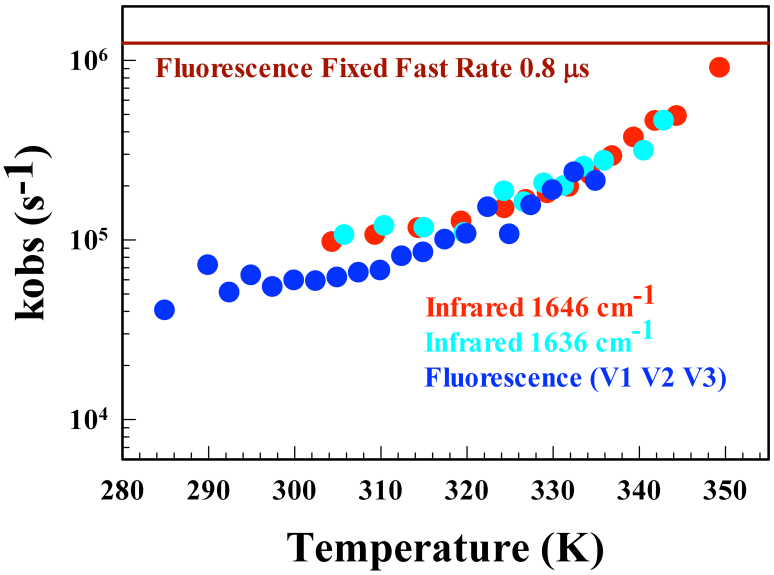


Figure 5-1. *Comparison of relaxation rates from infrared and fluorescencet T-jump measurements.*

performed in a sequential way from equilibrium to kinetics. Global analysis of the entire thermal unfolding experiments, including that of the unfolding measurements by fluorescence and infrared, to Mean Field Model had estimated the thermodynamic parameters of the model. Using this, probability distributions as a function of nativeness could be obtained for initial and final temperature of a temperature jump measurement. The absolute value of the signal (<S>) as a function of the order parameter, nativeness (n), for every λ, for initial and final temperature of a temperature jump measurement could also be obtained from the above analysis. Decays were considered to diffusive and a diffusion parameter had to be defined at every temperature to represent the kinetics. The decays could then be simulated by the rate matrix method (See Materials and Methods). In the case of engrailed kinetics data, all the decays could be fit in this manner. Diffusion values obtained from this analysis were fast and were in the range between 10^6 – 10^8 s^{-1} between two extreme final temperatures measured. Using the Kramer's like equation to estimate the diffusion parameter did not yield good fit of the

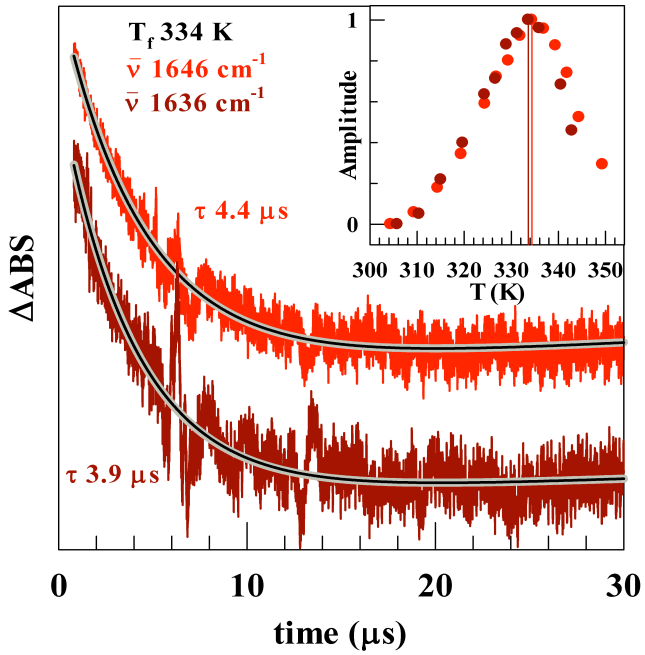


Figure 5-2. *Infrared decays measured at two wave numbers 1646 cm^{-1} and 1636 cm^{-1} at 334 K. Single exponentials fits are shown in grey and fits to mean field (1DFES) model are shown in black. In the inset, amplitude change for both the wave numbers are shown.*

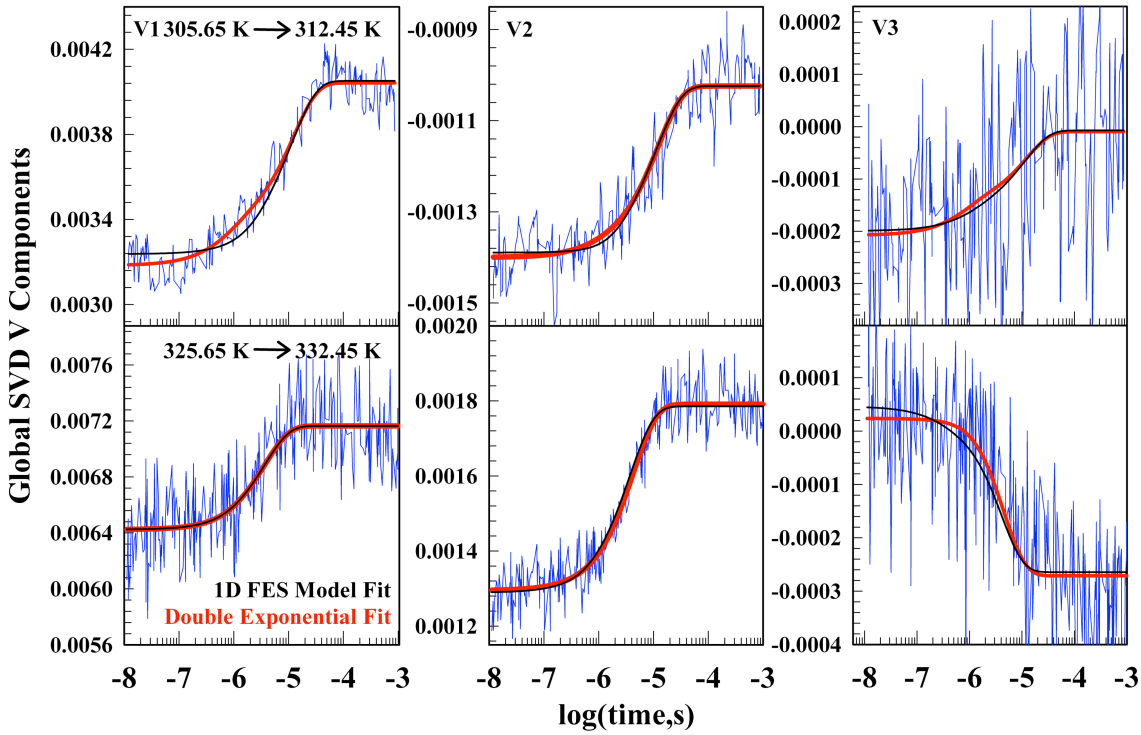


Figure 5-3. *Fluorescence T-jump decays for the three decay components (V) from the Global SVD analysis at two final temperatures are shown. Upper and lower panels represent three components at a particular final temperature of the t-jump respectively. Double exponentials fits are shown in red and fits to mean field (1DFES) model are shown in black.*

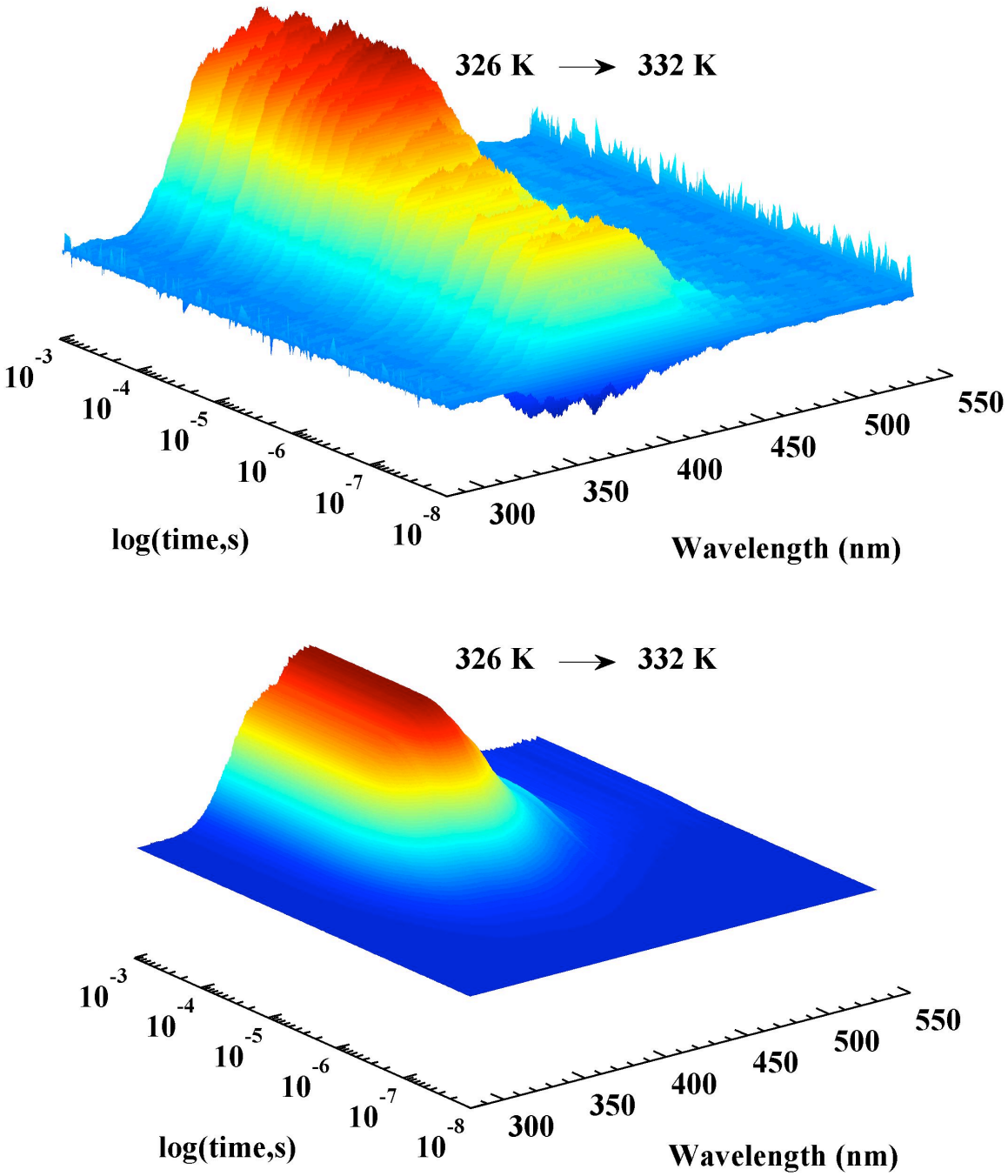


Figure 5-4. *Top panel shows the spectral decay from the experiment corresponding to a temperature jump of 326 K to 332 K. Bottom graph is the fit of this experimental data to 1D FES model.*

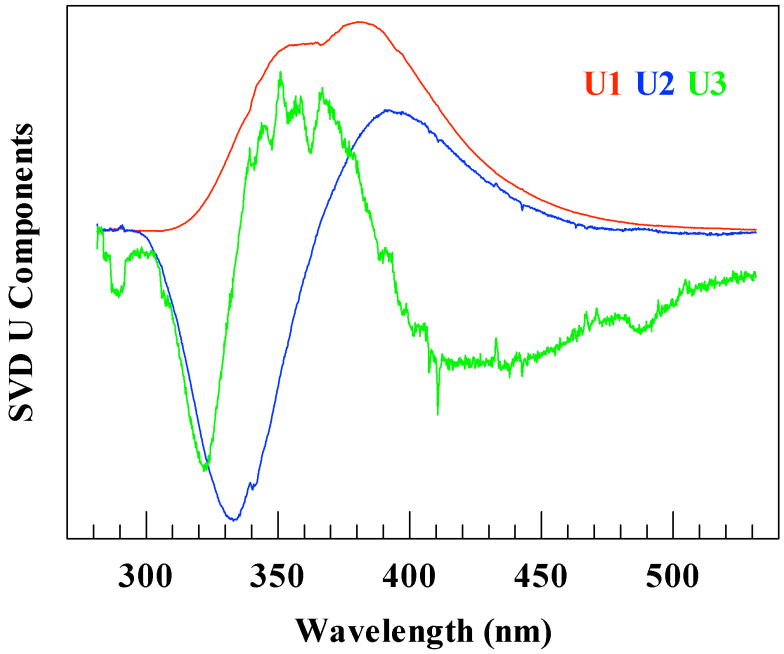


Figure 5-5. *Wavelength components (U) from the Global SVD analysis of all the fluorescence decays.*

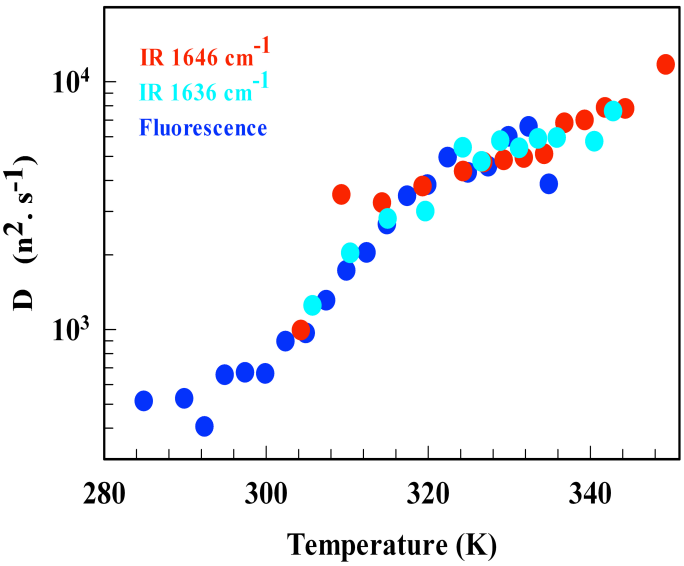


Figure 5-6. *Diffusion coefficient estimated from the fit of the decays to 1DFES model.*

Table 5-1. *Mean Field Model Fit Parameters*			
$\Delta C_{p,res}$ J.mol-1.K-1	$\Delta H_{loc,res}$ $kJ.mol^{-1}$	$\Delta H_{nonloc,res}$ $kJ.mol^{-1}$	$\beta(T_0)$ $kJ.mol^{-1}$ (K)
0	4.36	3.46	1.3 (326.2)

decays, as the diffusion parameter did not vary as represented by that equation. Thus, diffusion parameters were allowed to freely vary with temperature and for each probe. SVD analysis of the fit spectral decay revealed the same three significant components as observed in the experimental equilibrium and kinetic data. They contributed to ~ 16 % and 0.7 % with respect to the overall fluorescence signal.

5.4 Discussion

If EnHD folded by a two state fashion, only one set of rates should be observed for as many number of probes measured and the maximum amplitude change from each of the probes should occur at the same temperature. If the folding of EnHD occurred in a three-state fashion, one should observe two rates from all the probes and the maximum amplitude change for the slower phase should coincide for all the probes and for the fast phase as well. Estimated apparent barrier height to folding near characteristic temperature for EnHD was ~ 0.47 RT. This falls within downhill folding regime. If the folding mechanism were to be downhill, there should have been differences in T_m between different probes but the rate of exchange should have been the same between different probes. What was observed for engrailed was single exponential kinetics for 2 IR frequencies and double exponential for the fluorescence. The slow rate from fluorescence agreed with that of the rates from the infrared. The corresponding amplitudes showed

differences the melting temperature. This kinetic phase could be taken to exert a downhill behaviour. BBL showed such differences in melting temperature and same rates between two probes. EnHD kinetics is complicated than BBL in the sense that more components could be revealed by fluorescence kinetics and also a faster phase.

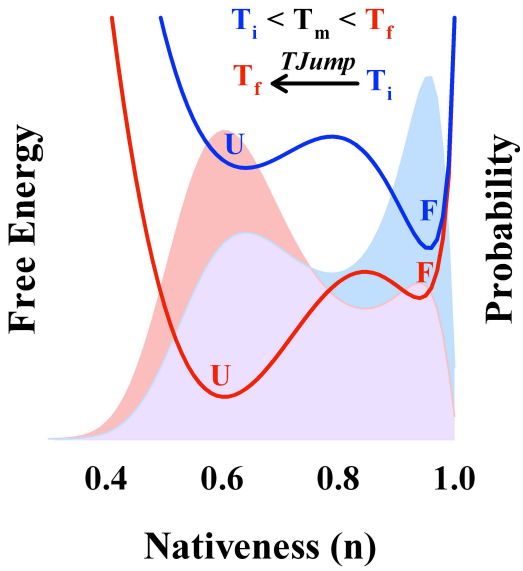


Figure 5-7. *Population redistribution upon introducing a temperature jump from an initial temperature less than the characteristic temperature to a final temperature greater than characteristic temperature.*

The origin of the faster phase was attributed to the presence of a marginal (low) barrier in one of the WW domains[25,96]. It was explained that proteins diffusing through the transition region, because of the low barrier, resulted in the observation of signals from the transition region in the form of a fast kinetic phase. This was called 'molecular phase'. In our case, this fast kinetic phase was observed only in the fluorescence. Fluorescence amplitude of EnHD had maximum signal change at low temperature in comparison to that of infrared. In other words, there would not be any signal change leading to relaxation for engrailed (See Chapter 3 – Two-State Comparison), at low temperatures by infrared temperature jump and at high temperatures by fluorescence T-jump. As the characteristic temperature of unfolding for EnHD was ~ 326 K, at the temperatures when infrared t-jump produced significant signal change, most of the

protein would have diffused to the unfolded side of the free energy surface and started to continuously unfold. But at the temperatures where fluorescence produced significant signal change, substantial amount of protein had to pass through the transition region. This could probably be why the molecular phase was observed only in the case of fluorescence kinetics and not by infrared kinetics. It was also evident when the fluorescence decays became single exponential decays at high temperatures measured and agreed with of single exponential behaviour of infrared.

This transition in the characteristic was also seen in the trend of diffusion coefficient calculated at each temperature. As expected, diffusion coefficient increased with temperature. But, the rate at which the diffusion coefficient was increasing was more at the temperatures where fluorescence decays were measured than the temperatures at which infrared kinetics were measured. The diffusion values at the highest measured temperatures for the fluorescence reached same rate of increase as that of infrared.

The conventional two or three-state analysis cannot account for the differences in the amplitude[23] between different spectroscopic probes. Also, in the case of EnHD, rates have to be analyzed by a two-state scheme for the infrared and three-state scheme for the fluorescence. Both schemes would not capture the underlying mechanism.

All these results and explanations are convincing enough to say that engrailed folds in a downhill manner and the origin of fast phase can be ascribed to the presence of a marginal barrier.

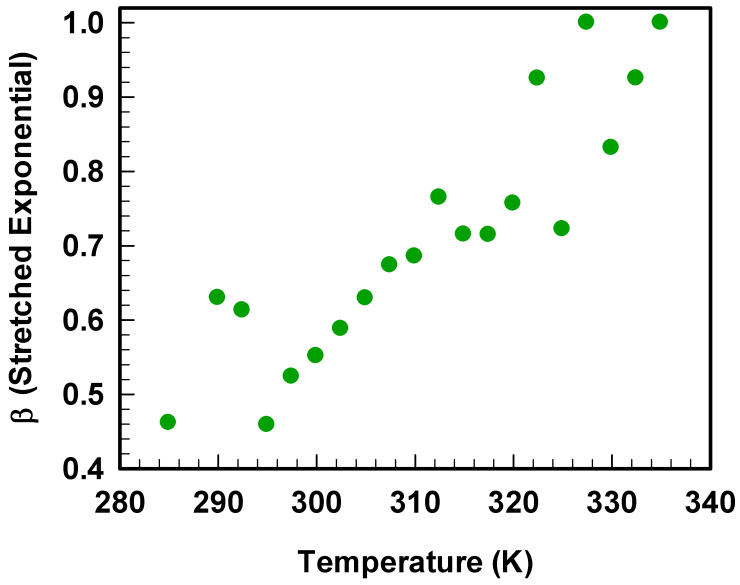


Figure 5-8. *Stretched exponential exponent by fitting the fluorescence temperature jump decays to a stretched exponential.*

Chapter 6

Exploring the folding mechanism of engrailed homeodomain by single molecule FRET spectroscopy

6.1 Abstract

Engrailed homeodomain has been shown to exhibit complex ultrafast folding kinetics with non-exponential decays and probe-dependent amplitudes. The folding barrier estimated near characteristic temperature for this protein suggests that the protein folds by a downhill mechanism and the additional faster phase observed can be claimed to be the molecular phase coming from the transition ensemble. Here, we explore the conformational distribution and the transition path of the engrailed homeodomain using single molecule FRET spectroscopy near mid-denaturing conditions. We also reconstruct energy landscape of the engrailed homeodomain by implementing the maximum likelihood method to analyze the photon arrival times using one-dimensional free energy surface model and estimate barrier heights from this analysis.

6.2 Introduction

Single molecule FRET spectroscopy has been used to study protein folding for more than a decade. smFRET studies on two-state proteins, that is the proteins that fold over a large free energy barrier, show two well-defined peaks separated from one another even at mid-denaturing conditions. Increasing or decreasing the denaturant concentration would increase or decrease the folded/unfolded subpopulations, but wouldn't result in the movement of subpopulations. In other words, the conformational distribution in between the folded and unfolded populations cannot be even detected by single molecule measurements for these proteins.

A small protein, BBL, has been shown to fold in a downhill fashion (barrierless) by a number of experimental techniques. BBL is a fast folding protein with a relaxation time of ~ 20 μs at room temperature[23]. Performing smFRET experiment at this condition wouldn't help resolve the conformational distribution because of fast folding kinetics and the limitation on the number of photons that can be obtained within this time as it wouldn't not give enough statistics to represent the conformational distribution (FRET) in terms of a histogram. But, the relaxation time of this protein slows down to ~ 120 μs at low temperature (279 K). This allowed the use of 50 μs binning time to analyze the photon trajectory. smFRET histogram obtained by analyzing the photon trajectory using this binning time clearly showed uni-model conformational distribution for this protein and any increase/decrease in the denaturant concentration resulted in the movement of the uni-model distribution in the appropriate direction. Using 50 μs binning time has been the limit while analyzing smFRET trajectories of a protein folding process. Trolox-

Cysteamine was used in the above smFRET measurement in order to protect the fluorophores used (Alexa 488/Alexa 594) in the above measurement from photobleaching and photoblinking[35,37]. smFRET measurements of a downhill folding protein would result in exploring the intermediate conformational distribution of the protein.

In this current work, we study an ultrafast folding protein, engrailed homeodomain, by smFRET spectroscopy. Engrailed homedomain has been shown to fold over a very small barrier of ~ 0.5 RT near mid-denaturing (T_m) condition, falling within the downhill folding limit but not globally downhill (barrierless). Though the calculation of barrier height propose a downhill folding mechanism for this protein, multiple probe kinetic studies on this protein has revealed an extremely complex fast folding kinetics for a protein of this size. Kinetic studies monitoring five properties of engrailed using two spectroscopic probes have shown non-exponential kinetics, differences in the kinetics between different spectroscopic probes (exponential in one probe/non-exponential in another probe) and probe-dependent amplitudes between five processes. In the case of BBL, multiple probe kinetics studies resulted in the same kinetics between different two probes and probe-dependent amplitudes[23]. Thus, studying such an extremely complex fast folding protein, engrailed, with a very small barrier height to folding, by smFRET could be very interesting and could be different from BBL and other two-state proteins. It would also help explore the complex conformational distribution of engrailed. It is also of the current interest to actually extract the energy landscape of engrailed from the single molecule experiments itself.

6.3 Results and Discussion

6.3.1 Fluorescence T-Jump Kinetics Near Chemical Denaturation Midpoint

Slow relaxation rate of engrailed homeodomain obtained at 285 K was ~ 25 μs, while keeping the fast rate to be constant at 0.8 μs. Performing smFRET folding measurement at this condition is going to be difficult because of the fast rate and the limitation on the binning time that can be used to analyzed the single molecule data. Thus, smFRET measurement had to be performed near to the apparent chemical denaturation midpoints, as it would slow down the kinetics of the protein.

Two-state analysis of chemical denaturation curves by steady state fluorescence at a series of temperatures between 278 K and 308 K showed the denaturation midpoints ranged from 3.5 M at 278 K to 1.5 M at 308 K. A series of T-Jump fluorescence kinetic measurements were performed on unlabeled engrailed homeodomain near the apparent chemical denaturation midpoints obtained from steady-state fluorescence at 281 K, 288 K, 298 K and 308 K. About 4-7 K temperature-jumps were produced on the sample. SVD analysis of the spectral relaxation at each temperature revealed two significant components, one corresponding to the average fluorescence signal (QY) and another corresponding to the spectral shift upon unfolding. This is in good agreement with the first two components resulting from a SVD analysis performed on the ensemble chemical denaturation measurements of EnHD by fluorescence. Ensemble measurements could also reveal a third component that couldn't be resolved by kinetics measurement. Some decays from these measurements looked double exponential, whereas others were stretched. At 288 K, the first component relaxation was double exponential and the

second component relaxation was single-exponential. Though these results looked slightly complex, decays were uniformly fit to a double exponential at each temperature and results are tabulated. The major relaxation phase at 281 K was ~ 0.5 ms, where as the major relaxation rate at 308 K was ~ 33 µs. The fast rates obtained at all these temperatures near mid-denaturing conditions were less than 5 µs. These results provide initial information needed to perform the single molecule measurements on a fluorescently labeled engrailed homeodomain at these conditions. EnHD was labeled with Alexa 488 as donor and Alexa 594 as acceptor.

Table 6-1 Fluorescence T-Jump Kinetics Near Cm

Double Exponential Fit Parameters			
T[a](K)	[UREA][b](M)	$\tau 1(\mu s)$	$\tau 2(\mu s)$
281	3.63	518	2.88
288	3.15	165	4.5
298	2.02	46.7	3.45
308	2.08	32.9	1.38

[a] Final Temperature of T-Jump
[b] UREA Concentration measured by Refractive Index

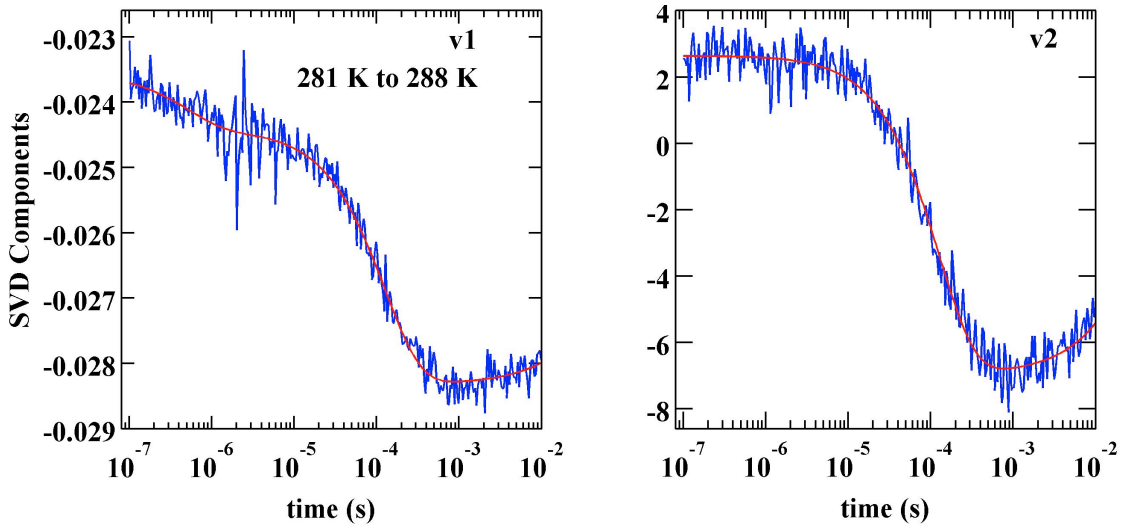


Figure 6-1. *Fluorescence T-Jump Decays near chemical denaturation midpoint at the final temperature of t-jump of 288 K. The two components shown here are the two SVD V component decays at that temperature. Experimental data are shown in blue and the double exponential fits are shown in red.*

6.3.2 Single Molecule FRET Measurements Near Chemical Denaturation Midpoint

6.3.2.1 Burst Identification by Clustering

A burst is a stream of photons in quick succession. A bin is the time-interval in which the photon trajectories are equally divided throughout the length of the trajectory. Photon trajectories (both donor and acceptor) have traditionally been analyzed by binning the trajectories and such trajectories can be analyzed for different time-bins/time-intervals. In order to facilitate the identification and selection of photon bursts in a trajectory, a k-means clustering procedure has been introduced. K-means clustering procedure selects for a particular photon burst at the appropriate time-length. In other words, every burst selected by this procedure has a unique time-length. The procedure would also retain the photon arrival time information from the trajectory and this eliminates the needs for binning the trajectory, as it would modify the trajectory before the analysis. If the dynamics of the process under study occur on the same time-scale of binning or faster, it would obscure the dynamics. smFRET trajectories of engrailed homeodomain were analyzed by this clustering procedure and ~ 500 bursts were selected from every 10,000 photons. In this 500 photon bursts, some photon bursts/clusters would correspond to the signal from the molecule and others are background photons without information from the molecule.

The choice of number of clusters/bursts that had to be identified in a given data depends largely on the data and an optimal choice of number of clusters had to result after analyzing the trajectory for a series of values of number of clusters. A high value for the number of clusters might chop many ‘real’ photo bursts into several bursts and a low

value would join many photon bursts together through out the length of the data. This can be easily identified from the average time-length of all the clusters or photon bursts. Knowing the relaxation rate of the protein beforehand could also help if the average time-length of the cluster from the clustering analysis makes sense or not.

Though single molecule experiments of engrailed were performed at 281 K, 288 K, 298 K and 308 K at the corresponding mid-denaturing conditions, only the experimental result at 288 K was considered. This was because the relaxation rate at 281 K was too slow for the free diffusion smFRET experiments and the ones at 298 K and 308 K can be extremely fast in order to obtain the right conformational distribution of the molecule at that condition. This is because when the dynamics are extremely fast, there could be two molecules diffusing in quick succession or the same molecule can rise to the number of photons corresponding to two molecules quickly and the photon burst cannot be separated. Slow relaxation rate at 288 K was ~ 127 μs and hence was suitable to perform smFRET experiments at that condition and to obtain the right information from these results. The choice of experimental condition considered was very much evident in the single molecule results obtained at those conditions.

6.3.2.2 Selection of Bursts

Selection of photon bursts corresponding to the FRET signal of the molecule involves the following procedure:

i) Clustering procedure divides a photon trajectory into ‘n’ clusters/photon bursts of unique time-length. From this, bursts corresponding to the signal from the molecule had to be selected for and background/non-informational bursts had to be discarded. For

each cluster/burst identified, photon count rates were calculated as the ratio of total number of photons present in the cluster to time-length of that particular cluster. Time-length of a cluster is the time spent by the molecule in the confocal volume leading to a photon burst. A threshold for the count rate was set one standard deviation above the average photon count rate. Only those bursts that have the photon count rates above the threshold were taken as the photon bursts corresponding to molecule.

ii) Once the photon burst corresponding to the molecule has been selected, the photon bursts have to be classified as the bursts corresponding to the FRET of the molecule and the bursts coming solely from the donor molecules contributing to zero peaks. (FRET value of a particular cluster/photon burst was calculated as the ratio of number of acceptor photons to total number of photons present in that particular cluster.) When the number of acceptor photons present in the bursts is represented as Poisson distribution, the bursts/clusters corresponding to < 1 % probability obtained from the distribution are classified as good bursts contributing to the FRET of the molecule as this < 1 % burst would have more acceptor photons for the duration of the burst or would have good acceptor count rates and extremely likely to be ones contributing to the FRET. The rest of the bursts are classified as zero peaks.

Though the selected 'good' photon bursts should contribute to the FRET value from the molecule, it could have other experimental artifacts such as acceptor blinking. This can be primarily identified from the inter acceptor time-lengths. An extremely high value for the inter acceptor time-length when compared to the average value can indicate acceptor blinking and those photon bursts have to be discarded. At this point, a threshold could be set for the number of photons present in a photon burst in order to select that burst and

this was basically done to have only good quality data to be used for any further analysis. In the case of smFRET experiments performed on engrailed near mid-denaturing conditions at 288 K, about 10,000 photon bursts were finally selected according to this procedure.

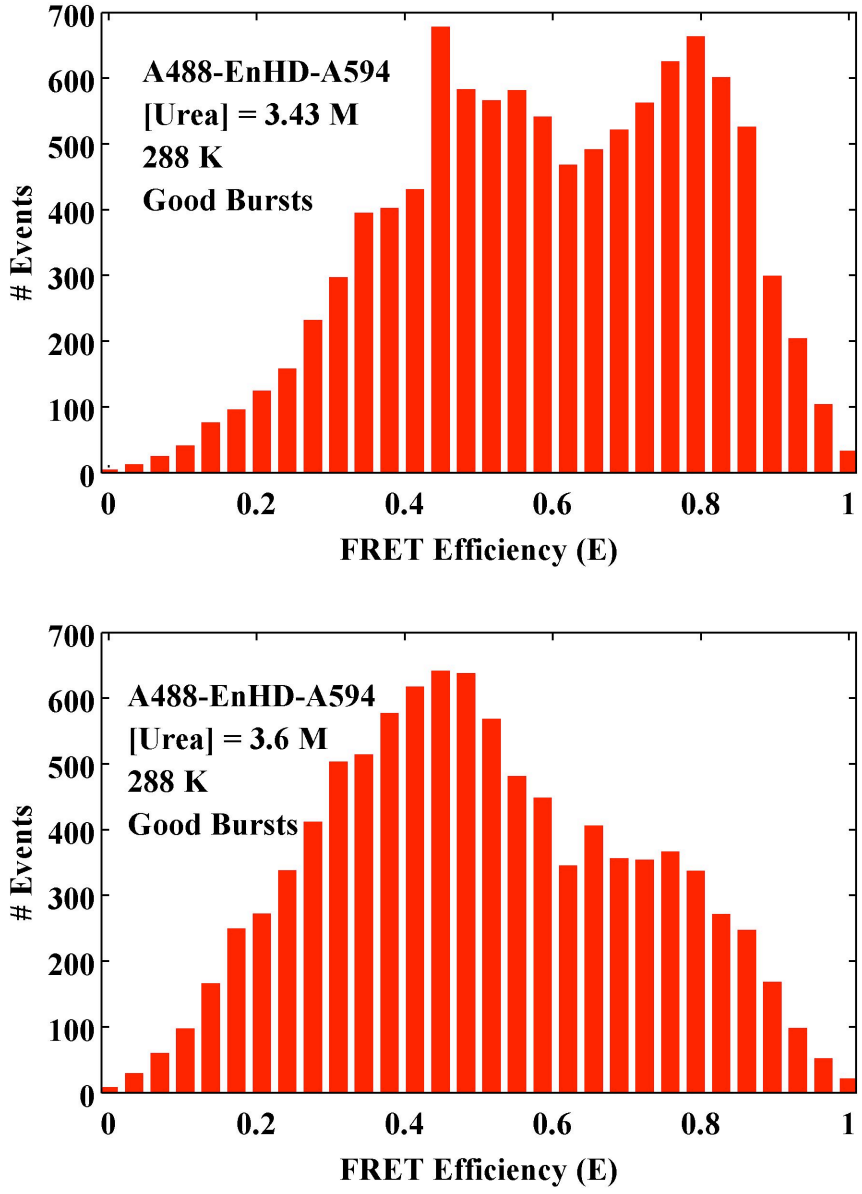


Figure 6-2. *Histograms representing the results from single molecule FRET experiments performed on EnHD near mid-denaturing conditions at 288 K. The events represented here are the good bursts selected after filtering for acceptor blinking as described in the text. Experimental conditions are indicated in the figure itself.*

6.3.2.3 Analysis of Photon Arrival Times of Bursts Selected Using 1D FES Model by Maximum Likelihood Method

Results from Histogram: Two smFRET experiments were performed near mid-denaturing conditions at 288 K at urea concentrations of 3.43 M and 3.6 M respectively and the results from these experiments after the selecting for the burst by clustering analysis and filtering for the 'good bursts' were plotted as histograms. Histogram obtained from the experimental result at 3.43 M urea concentration showed two peaks that are broad and overlapping with each other, with the lower FRET peak broader than the higher fret peak. Higher fret and lower FRET peaks showed almost equal amount of sub-populations. In the case of the smFRET performed at 3.6 M urea concentration, lower FRET peak had more population when compared to the higher FRET peak and still both the peaks were broad and overlapping. When the results from two experiments are compared, both the folded (higher FRET peak) and unfolded (lower FRET peak) subpopulations of engrailed moved towards unfolded side/lower fret values and also resulted in the disappearance of folded populations and leading to an increase in the unfolded populations when the concentration of denaturant (urea) was slightly increased. This was quite intriguing in the sense that both the scenarios both the folded and unfolded sub-populations heavily overlapped. An estimation of apparent chemical denaturation mid-point from the bulk FRET experiments at 288 K from the two-state analysis was 3.9 M urea. But, the results from the single molecule experiments showed the mid-denaturing condition happened to fall between 3.43 M and 3.6 M urea concentration for engrailed homeodomain. This result, together with the broadness of peaks from histograms or the

complete spread of sub-populations throughout the length of the FRET values, indicated a complex folding scenario, yet a bimodal distribution confirmed the presence of a (small) barrier. The single molecule results, indeed clearly visualized the transition region.

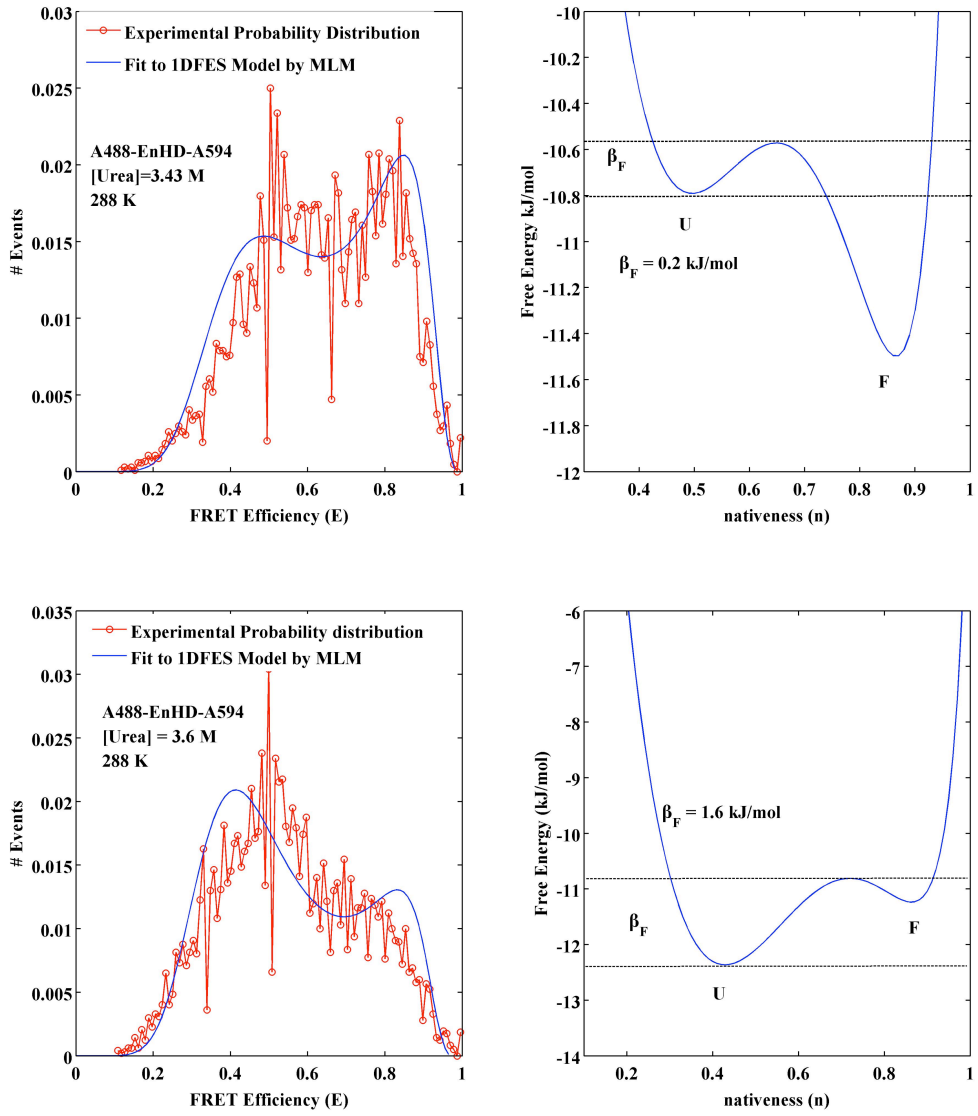


Figure 6-3. *Results from the analysis of smFRET measurements by 1DFES/MF Model Model using Maximum likelihood method to analyze photon trajectories.*

The results were analyzed with a model-free approach using a statistical mechanical model (Mean Field Model). The one dimensional free energy surface model was used in conjunction with maximum likelihood analysis of photon arrival times (See Materials and Methods). Likelihood analysis[74] was used analyze the trajectory, one photon burst after another, directly from the arrival times of photon and probability of the color of the photons (donor or acceptor). The combined analysis estimated the best parameters for the 1D FES[68,72] model by maximizing the combined log-likelihood value calculated for all the photon bursts taken for the analysis. Model parameters are tabulated (Table 6-2) and the free energy surface produced by the model are shown in the figure.

***Table 6-2**. Mean Field Model/1DFES Parameter Model Parameters*					
[UREA] M	ΔH_{res} $kJ.mol^{-1}.res^{-1}$	$\kappa_{\Delta H}$	$FRET_0$	$\Delta FRET$	D(288 K) $n^2.s^{-1}$
3.43	5.49	1.79	0	0.98	256
3.6	5.42	1.85	0	0.96	226

where n represent the nativeness interval in the unit.

The barrier height estimated from the resulting analysis was 0.22 kJ/mol (~ 0.1 RT) at 3.43 M urea concentration and 1.6 kJ/mol (~ 0.65 RT) at 3.6 M urea concentration. The estimated barrier heights near mid-denaturing conditions were less than 1 RT and thus, the over-all folding behavior of engrailed homeodomain, though complex, can still be explained by a downhill mechanism and would also explain the results shown from the histogram. Barrier height estimated from bulk thermal denaturation experiments was also < 1 RT near characteristic temperature and would confirm to the conclusions made from the single molecule experiments.

Conclusions

A conventional two/three state analysis assumes (a) large barrier(s) (> 4RT) between state(s), whereas a downhill folding implies lesser barrier height between states (< 2RT). Thus, experimental results will be discussed from both these contexts.

1) Multiple probe thermal and chemical unfolding measurements resulting in broad non-coincidental unfolding curves between different spectroscopic probes leading to differences in melting temperature or apparent chemical denaturation midpoint is a clear signature of the presence of downhill folding mechanism. Large barrier(s) would anyways produce overlapping unfolding transitions by different probes. These experiments, however, cannot resolve multiple marginal barriers within downhill limit. Yet, global analysis of all the thermal unfolding curves including that of the thermogram from differential scanning calorimetry by a free energy surface statistical mechanical model can estimate the barrier height near characteristic temperature. For, engrailed it was estimated to be 0.5 RT and thus the overall thermal unfolding behavior falls within the downhill regime.

2) Analysis of double perturbation experiment (represented as thermal unfolding curves) can produce complex results, if the folding mechanism deviates from two-state. In the case of previous knowledge about inexistence of additional states, this can be set as a criterion to identify downhill folding. But, for more than two-states, it is complicated to

interpret. Additional results, such as differences in C_m as it has been observed in engrailed, would mark a downhill behavior, but cannot resolve multi-state behaviors.

3) Multiple probe kinetic measurements can conclusively say, a given kinetic phase that is commonly observed in all the probes, if that phase results in a downhill or conventional two-state behavior, based on the extent of overlap of kinetic amplitudes. In the case of engrailed, infrared T-jump measurement resulted in single exponential decays and fluorescence T-jump measurement produced non-exponential decays. Non-exponential decays fit very well the double exponential. Probe-dependent kinetic amplitudes for the slow rate imply a downhill mechanism for that kinetic phase. The observation of faster rate can be interpreted as the proteins diffusing through the transition region, because of the low barrier, resulted in the observation of signals from the transition region in the form of a fast kinetic phase. Thus, all the kinetic decays and thermal unfolding measurements were globally fit to MF model. Even if additional faster phases were observed on this timescale in both probes, data could still be analyzed similarly.

4) Single molecule measurements are the only way to visualize conformational distributions and the results (population distribution) from these measurements can be directly used to estimate the barrier height of engrailed homeodomain. A conventional three-state folder must show three distinct peaks at a particular experimental condition. A protein with two marginal barriers can also produce three peaks, depending upon the denaturant stress, but would also result in the movement of peaks depending upon the

denaturant concentration. In the case of engrailed, T-jump kinetic measurements revealed a faster phase of 0.5 μs and a slow phase of ~ 127 μs near C_m at 288 K. Only the slow phase can be resolved by smFRET measurements. smFRET measurements near C_m at 288 K, ulitizing the methods available to produce high temporal resolution, revealed a bimodal distribution for this protein. This could imply a re-equilibriation of conformations at these timescales or the presence of a significant population at the top of the free enery barrier. To resolve this problem, we utilize the analysis of maxiumum likelihood method about the photon arrival times in combination with the free energy surface statistical mechanical model that was used to analyze the thermodynamics and kinetics of this protein. The advantage of this method is that it permits to expand the temporal resolution of the analysis of photon arrival times (on the order one micro second) in the cases where accumulation of a lot of photons in a defined timescale would be necessary to analyze these experiments by conventional methods. This analysis resulted in the estimation of free energy surface of engrailed homeodomain and the conformational dynamics of the protein resulting from this analysis confirmed the folding of this protein to be downhill and controlled by the crossing of the constant marginal (small) thermodynamic barrier.

This research work shows different experimental folding studies ranging from bulk thermodynamic measurements to ultra fast kinetics to single molecule measurements and how the results from different measurements can be related and how the results from these different experiments provide necesary details in order to conclude the folding of engrailed homeodomain to be downhill and not conventional three-state. This work also

talks about the use of statistical mechanical model to analyze different experimental results and how the global analysis of complicated experimental data of this protein by this model help correcting the qualitative interpretations that are made by the other research groups about the folding mechanism of this protein.

Conclusiones

Los resultados experimentales obtenidos mediante múltiples técnicas termodinámicas, cinéticas y de molécula única has sido interpretados y analizados desde los puntos de vista alternativos que ofrecen los modelos bioquímicos convencionales de plegamiento proteico que asumen la existencia de grandes parreras de energía libre separando dos (nativo y desplegado) o tres (nativo, intermediario y desplegado) estados de plegamiento, y por otro lado el análisis del plegamiento en térmidos de modelos estadísticos de superficie de energía libre en los que las barreras de energía libre son modulables, llevanod incluso a la aparación de plegamiento de tipo *downhill*.

1) La obtención de curvas de desplegamiento en equilibrio no coincidentes según la sonda espectroscópica utilizada y, por tanto, la determinación de diferencias en los parámetros termodinámicos (punto medio de desnaturalización, T_m o C_m, y cambio en entalpía de desnaturalización, ΔH) obtenidos con cada sonda es una clara señal de existencia de un mecanismo de tipo downhill para el homeodominio engrailed, tanto en el contexto de la desnaturalización térmica como química. Estos resultados son suficientes para poder rechazar la existencia de plegamiento a través de barrieras de energía libre altas, dado que en este caso se deben observar transiciones de desplegamiento superpuestas para todas las sondas espectroscópicas utilizadas. Sin embargo, estos experimentos por sí solos no pueden discernir entre la existencia de barreras marginales múltiples de plegamiento o un modelo típico de tipo downhill. Sin embargo, un análisis cuantitativo global de todas las curvas de desplegamiento térmico, incluyendo a su vez los termogramas obtenidos mediante calorimetría diferencial de barrido, utilizando un modelo mecánico estadístico de superficie de energía libre demuestra que se pueden

explicar todas las observaciones pexerimentales con un modelo tipo downhill en el que la barrera de energía libre es de sólo 0.5 RT, es decir significativamente más baja que la energía térmica.

2) El análisis de los experimentos de desnaturalización en equilibrio utilizando doble perturbación (desplegamiento térmico y químico) debería dar lugar a resultados complejos cuando el mecanismo de plegamiento en equilibrio se desvía significativamente del tipo dos-estados. Dicho análisis, en un caso donde no haya conocimiento previo sobre la existencia de estados adicionales, puede tomarse como un criterio en la identificación del plegamiento downhill. Aun así, la dificultad en la interpretación de aquellos casos en los que se trata de distinguir entre un modelo downhill o la presencia de varias barreras de energía libre, siendo ambos modelos coincidentes en su mayor complejidad sobre el plegamiento dos-estados, se necesita la recopilación de resultados que ofrezcan pruebas diagnósticas adicionales. En el caso del homeodominio engrailed se aportan evidencias adicionales, como la observación de cambios en los parámetros termodinámicos no sólo de desnaturalización térmica, sino también química (cambios en C_m y m) que apoyan su identificación como plegamiento tipo downhill, aunque no es todavía suficiente evidencia para demostrar esto de una manera concluyente.

3) Las medidas cinéticas ultrarrápidas usando salto de temperatura inducido en nanosegundos combinado con medidas de infrarrojo y fluorescencia pueden concluir si una fase cinética observada con todas las sondas define un plegamiento downhill o dos estados convencional en función de un criterio de superposición de las amplitudes cinéticas. En el caso del homeodominio engrailed, las medidas de salto de temperatura en

el infrarrojo resultaron en decaimientos exponenciales simples mientras que medidas equivalenters de fluorescencia produjeron decaimientos no exponenciales con buenos ajustes a una función doble exponencial. El que las amplitudes cinéticas dependan de la sonda espectróscopica son señal diagnóstica de la existencia de mecanismos de plegamiento downhill. La observación de una fase más rápida adicional con algunas sondas, sin embargo, puede interpretarse como que dicha fase rápida se produzca debido a la formación de un intermediario de plegamiento a a la fase molecular cuando la barreras de energía libre es marginales y por lo tanto existe una población significativa en el tope de la barrera. El análisis de los resultados obtenidos mediante fluorescencia revela un proceso complejo, en el que se pueden discernir tres componentes distintos dentro del análisis SVD de los datos, que se manifiestan en diferencias en las amplitudes y en los decaimientos cinéticos. Finalmente, el análisis global de todos los experimentos cinéticos multi-sonda mediante un modelo de mecánica estadística de superficie de energía libre revela que es posible explicar todas estas observaciones global y cuantitativamente utilizando las misma superficie de energía libre obtenida del análisis multisonda realizado previamente en equilibrio. Es decir, este análisis confirma las conclusiones extraídas en los apartados previos y aporta evidencia cinética de plegamiento tipo downhill en el homeodominio engrailed.

4) Las medidas de fluorescencia en moléculas únicas es posiblemente la única manera existente actualemente de visualizar las distribuciones conformacionales de una proteínas y de esta manera poder estimar directamente la altura de la barrera de plegamiento del homeodominio engrailed. Con esta técnica, un dominio quer exhiba desplegamiento químico de tres estados, como se ha postulado anteriormente para este

dominio, debería mostrar tres picos diferentes a distintas concentraciones de desnaturalizante correspondientes a las poblaciones de los tres estados. Una proteína en la que las barreras separando a estos estados sean marginales puede también dar lugar a tres picos en función del estrés de desnaturalización, sin embargo, en este caso los picos deberían estar parcialmente solapados y cambiar sus propiedades (su valor de FRET) en función de la concentración de desnaturalizante. En el caso del homeodominio engrailed, las medidas cinéticas de salto de temperatura mediante FRET usando las mismas sondas a utilizar en las medidas de molécula única revelaron una fase rápida de 0.5μs y una fase lenta de ~ 127μs en condiciones cercanas a la C_m a 288K. En medidas de FRET en moléculas únicas solo pudo ser resuelta la fase lenta debido a la falta de resolución temporal. Las medidas smFRET cerca del C_m utilizando los métodos disponibles de mayor resolución temporal muestran una distribución bimodal muy ancha que es indicativa ya sea de re-equilibrado conformacional durante la escala de tiempos de la medida, o de una población significativa del tope de la barrera de energía libre. Para resolver este problema utilizamos un análisis de probabilidad máxima sobre los tiempos de llegada de fotones individuales combinado con el modelo mecánico estadístico de superficie de energía libre usado anteriormente parta analizar la termodinámica y cinética de plegamiento de esta proteína. La ventaja de este método es que permite expandir la resolución temporal al depender en este caso de los tiempos de llegada de fotones individuales (en el orden de 1 microsegundo) en vez de los tiempos necesarios para acumular conjuntos de varias decenas de fotones que se necesitan en el análisis convencional de estos experimentos. El resultado de este análisis resultó en la estimación de la superficie de energía libre del homeodominio engrailed así como de la dinámica

conformacional de la proteína, los cuales confirmaron que el plegamiento de esta proteína es de tipo downhill y controlado por el cruce de una barrera marginal siempre menor a la energía térmica.

Este trabajo de investigación presenta la realización de diferentes estudios experimentales del plegamiento de proteínas que incluyen medidas termodinámicas en bulk utilizando multiples sondas espectroscópicas, cinética ultrarrápida también multi-sonda y de espectroscopía de fluorescencia en moléculas únicas, y como estos experimentos aportan las pistas y pruebas diagnósticas necesarias para concluir que el plegamiento del homeodominio engrailed es del tipo downhill y no del tipo convencional de tres estados. Este trabajo ejemplifica también como la utilización de un modelo mecánico estadístico de superficie de energía libre para analizar la gran batería de datos experimentales obtenidos y que a su vez presentan gran complejidad aparente, puede ser determinante para mostrar la compatibilidad global de estos datos con los varios modelos de plegamiento, y poder así corregir las interpretaciones cualitativas hechas anteriormente por otros grupos de investigación sobre el plegamiento de esta proteína.

Bibliography

(1) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Proteins: Structure, Function, and Bioinformatics 1995, 21, 167.

(2) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. Science 2002, 298, 2191.

(3) Naganathan, A. N.; Sanchez-Ruiz, J. M.; Muñoz, V. Journal of the American Chemical Society 2005, 127, 17970.

(4) Fung, A.; Li, P.; Godoy-Ruiz, R.; Sanchez-Ruiz, J. M.; Muñoz, V. Journal of the American Chemical Society 2008, 130, 7489.

(5) Sadqi, M.; Fushman, D.; Muñoz, V. Nature 2006, 442, 317.

(6) Naganathan, A. N.; Perez-Jimenez, R.; Sanchez-Ruiz, J. M.; Muñoz, V. Biochemistry 2005, 44, 7435.

(7) Oliva, F. Y.; Muñoz, V. Journal of the American Chemical Society 2004, 126, 8596.

(8) Muñoz, V. International Journal of Quantum Chemistry 2002, 90, 1522.

(9) Callender, R. H.; Dyer, R. B.; Gilmanshin, R.; Woodruff, W. H. Annual review of physical chemistry 1998, 49, 173.

(10) Eaton, W. A.; Muñoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Annual review of biophysics and biomolecular structure 2000, 29, 327.

(11) Eaton, W. A.; Muñoz, V.; Thompson, P. A.; Chan, C. K.; Hofrichter, J. Current opinion in structural biology 1997, 7, 10.

(12) Gruebele, M. Annual review of physical chemistry 1999, 50, 485.

(13) Krieger, F.; Fierz, B.; Bieri, O.; Drewello, M.; Kiefhaber, T. Journal of molecular biology 2003, 332, 265.

(14) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. Journal of molecular biology 2002, 319, 19.

(15) Thompson, P. A.; Muñoz, V.; Jas, G. S.; Henry, E. R.; Eaton, W. A.; Hofrichter, J. The Journal of Physical Chemistry B 2000, 104, 378–389.

(16) Wang, T.; Du, D.; Gai, F. Chemical Physics Letters 2003, 370, 842–848.

(17) Werner, J. H.; Dyer, R. B.; Fesinmeyer, R. M.; Andersen, N. H. The Journal of Physical Chemistry B 2002, 106, 487–494.

(18) Xu, Y.; Oyola, R.; Gai, F. Journal of the American Chemical Society 2003, 125, 15388–15394.

(19) Muñoz, V.; Eaton, W. A.; Thompson, P. A.; Hofrichter, J. Nature 1997, 390, 196–199.

(20) Kubelka, J.; Hofrichter, J.; Eaton, W. A. Current Opinion in Structural Biology 2004, 14, 76–88.

(21) Thirumalai, D. J. Phys. I 1995, *5*, 1457-1467.

(22) Naganathan, A. N.; Muñoz, V. Journal of the American Chemical Society 2005, 127, 480–481.

(23) Li, P.; Oliva, F. Y.; Naganathan, A. N.; Muñoz, V. Proceedings of the National Academy of Sciences 2009, 106, 103–108.

(24) Yang, W. Y.; Gruebele, M. Biophysical Journal 2004, 87, 596–608.

(25) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. Proceedings of the National Academy of Sciences 2008, 105, 2369–2374.

(26) Forster, T. Modern Quantum Chemistry, Academic Press, New York 1965, 93–137.

(27) Stryer, L., and Haugland, R. P. Proc. Natl. Acad. Sci. USA 1967, 58, 719–726.

(28) Jia, Y.; Talaga, D. S.; Lau, W. L.; Lu, H. S. M.; DeGrado, W. F.; Hochstrasser, R. M. Chemical Physics 1999, 247, 69–83.

(29) Deniz, A. A.; Laurence, T. A.; Beligere, G. S.; Dahan, M.; Martin, A. B.; Chemla, D. S.; Dawson, P. E.; Schultz, P. G.; Weiss, S. Proceedings of the National Academy of Sciences 2000, 97, 5179–5184.

(30) Schuler, B.; Lipman, E. A.; Eaton, W. A. Nature 2002, 419, 743–747.

(31) Chung, H. S.; Gopich, I. V.; McHale, K.; Cellmer, T.; Louis, J. M.; Eaton, W. A. The Journal of Physical Chemistry A 2011, 115, 3642–3656.

(32) Boukobza, E.; Sonnenfeld, A.; Haran, G. The Journal of Physical Chemistry B 2001, 105, 12165–12170.

(33) Chung, H. S.; Louis, J. M.; Eaton, W. A. Biophysical Journal 2010, 98, 696–706.

(34) Chung, H. S.; Louis, J. M.; Eaton, W. A. Proceedings of the National Academy of Sciences 2009, 106, 11837–11844.

(35) Campos, L. A.; Liu, J.; Wang, X.; Ramanathan, R.; English, D. S.; Muñoz, V. Nature Methods 2011, 8, 143–146.

(36) Campos, L. A.; Sadqi, M.; Liu, J.; Wang, X.; English, D. S.; Muñoz, V. The Journal of Physical Chemistry B 2013, 117, 13120–13131.

(37) Liu, J.; Campos, L. A.; Cerminara, M.; Wang, X.; Ramanathan, R.; English, D. S.; Muñoz, V. Proceedings of the National Academy of Sciences 2012, 109, 179–184.

(38) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A. Science 2012, 335, 981–984.

(39) Gruschus, J. M.; Tsao, D. H. H.; Wang, L.-H.; Nirenberg, M.; Ferretti, J. A. Biochemistry 1997, 36, 5372–5380.

(40) Tsao, D. H. H.; Gruschus, J. M.; Wang, L.-H.; Nirenberg, M.; Ferretti, J. A. Biochemistry 1994, 33, 15053–15060.

(41) Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O. Protein Sci 1994, 3, 1779.

(42) Fraenkel, E.; Rould, M. A.; Chambers, K. A.; Pabo, C. O. J Mol Biol 1998, 284, 351.

(43) Religa, T. L. J Biomol NMR 2008, 40, 189.

(44) DiNardo, S. Cell 1985, 43, 59–69.

(45) Morgan, R. FEBS Letters 2006, 580, 2531–2533.

(46) Ades, S. E.; Sauer, R. T. Biochemistry 1994, 33, 9187.

(47) Ades, S. E.; Sauer, R. T. Biochemistry 1995, 34, 14601.

(48) Dyer, R. B. Curr Opin Struct Biol 2007, 17, 38.

(49) Gianni, S.; Guydosh, N. R.; Khan, F.; Caldas, T. D.; Mayor, U.; White, G. W.; DeMarco, M. L.; Daggett, V.; Fersht, A. R. Proc Natl Acad Sci U S A 2003, 100, 13286.

(50) Huang, F.; Settanni, G.; Fersht, A. R. Protein Eng Des Sel 2008, 21, 131.

(51) Mayor, U.; Grossmann, J. G.; Foster, N. W.; Freund, S. M.; Fersht, A. R. J Mol Biol 2003, 333, 977.

(52) Mayor, U.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M.; Alonso, D. O.; Daggett, V.; Fersht, A. R. Nature 2003, 421, 863.

(53) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. Proc Natl Acad Sci U S A 2000, 97, 13518.

(55) Neuweiler, H.; Banachewicz, W.; Fersht, A. R. Proc Natl Acad Sci U S A 2010, 107, 22106.

(56) Religa, T. L.; Johnson, C. M.; Vu, D. M.; Brewer, S. H.; Dyer, R. B.; Fersht, A. R. Proc Natl Acad Sci U S A 2007, 104, 9272.

(57) Religa, T. L.; Markson, J. S.; Mayor, U.; Freund, S. M.; Fersht, A. R. Nature 2005, 437, 1053.

(58) Makhatadze, G. I. In Current Protocols in Protein Science; John Wiley & Sons, Inc.: 2001.

(59) Makhatadze, G.I., Medvedkin, V.N., and Privalov, P.L. *Biopolymers, 1990,* 30, 1001.

(60) Dragan, A. I.; Li, Z.; Makeyeva, E. N.; Milgotina, E. I.; Liu, Y.; Crane-Robinson, C.; Privalov, P. L. Biochemistry 2006, 45, 141.

(61) Greenfield, N. J.; Fasman, G. D. Biopolymers 1969, 7, 595–610.

(62) Susi H, Byler DM. Methods Enzymol 1986, 130, 290.

(63) Byler, D. M.; Susi, H. Biopolymers 1986, 25, 469–487.

(64) Chirgadze, Y. N.; Fedorov, O. V.; Trushina, N. P. Biopolymers 1975, 14, 679–694.

(65) J. Zhang, R.E. Campbell, A.Y. Ting, R.Y. Tsien, Nat. Rev. Mol. Cell Biol. 2002, 3, 906.

(66) Johnson, M. L. Essential Numerical Computer Methods; Elsevier Science, 2010.

(67) Muñoz, V.; Sanchez-Ruiz, J. M. Proceedings of the National Academy of Sciences 2004, 101, 17646–17651.

(68) Naganathan, A. N.; Doshi, U.; Muñoz, V. Journal of the American Chemical Society 2007, 129, 5673–5682.

(69) Zwanzig R. Proc Natl Acad Sci USA. 1995;92:9801–9804.

(70) De Sancho, D.; Muñoz, V. Physical Chemistry Chemical Physics 2011, 13, 17030.

(71) Naganathan, A. N.; Perez-Jimenez, R.; Muñoz, V.; Sanchez-Ruiz, J. M. Physical Chemistry Chemical Physics 2011, 13, 17064.

(72) Lapidus LJ, Steinbach PJ, Eaton WA, Szabo A, Hofrichter J. J Phys Chem B. 2002;106:11628–11640.

(73) Gopich, I. V.; Szabo, A. Proceedings of the National Academy of Sciences 2012, 109, 7747–7752.

(74) Gopich, I. V.; Szabo, A. The Journal of Physical Chemistry B 2009, 113, 10965–10973.

(75) Beck, D. A.; Daggett, V. Methods 2004, 34, 112.

(76) Beck, D. A.; Daggett, V. Biophys J 2007, 93, 3382.

(77) Day, R.; Daggett, V. Adv Protein Chem 2003, 66, 373.

(78) DeMarco, M. L.; Alonso, D. O.; Daggett, V. J Mol Biol 2004, 341, 1109.

(79) Fersht, A. R.; Daggett, V. Cell 2002, 108, 573.

(80) McCully, M. E.; Beck, D. A.; Daggett, V. Biochemistry 2008, 47, 7079.

(81) McCully, M. E.; Beck, D. A.; Daggett, V. Proc Natl Acad Sci U S A 2012, 109, 17851.

(82) McCully, M. E.; Beck, D. A.; Fersht, A. R.; Daggett, V. Biophys J 2010, 99, 1628.

(83) Verma, A.; Wenzel, W. Biophys J 2009, 96, 3483.

(84) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. Science 2011, 334, 517.

(85) Greenfield, N. J.; Fasman, G. D. Biochemistry 1969, 8, 4108–4116.

(86) Manas, E. S.; Getahun, Z.; Wright, W. W.; DeGrado, W. F.; Vanderkooi, J. M. Journal of the American Chemical Society 2000, 122, 9883–9890.

(87) Purich, D. L. Enzyme Kinetics: Catalysis & Control: A Reference of Theory and Best-Practice Methods; Elsevier Science, 2010.

(88) Privalov, P. L.; Dragan, A. I. Biophysical Chemistry 2007, 126, 16–24.

(89) Sanchez-Ruiz, J. M. Annual Review of Physical Chemistry 2011, 62, 231–255.

(90) Pace, C. N.; Laurents, D. V. Biochemistry 1989, 28, 2520–2525.

(91) Robertson, A. D.; Murphy, K. P. Chemical Reviews 1997, 97, 1251–1268.

(92) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Proteins: Struct. Funct. Genet. 1995, 21, 167-195

(93) Socci, N. D.; Onuchic, J.; Wolynes, P. G. J. Chem. Phys. 1996, 104, 5860-5868

(94) Sabelko, J.; Ervin, J.; Gruebele, M. Proceedings of the National Academy of Sciences 1999, 96, 6031–6036.

(95) Hagen, S. J. Proteins: Structure, Function, and Bioinformatics 2002, 50, 1–4.

(96) Liu, F.; Nakaema, M.; Gruebele, M. The Journal of Chemical Physics 2009, 131, 195101.

(97) Bai, Y.; Nussinov, R. Protein Folding Protocols; Humana Press, 2007.

(98) Buchner, J. Protein folding handbook; Wiley-VCH, 2005.

(99) Callen, H. B. THERMODYNAMICS & AN INTRO. TO THERMOSTATISTICS; Wiley India Pvt. Limited, 2006.

(100) Dill, K. A.; Bromberg, S. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology; Garland Science, 2003.

(101) Fersht, A. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding; W. H. Freeman, 1999.

(102) Lakowicz, J. R. Principles of Fluorescence Spectroscopy; Springer, 2007.

(103) McQuarrie, D. A. Statistical Mechanics; University Science Books, 2000.

(104) Muñoz, V. Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches; Royal Society of Chemistry, 2008.

(105) Van Holde, K. K. E.; Johnson, W. C.; Ho, P. S. Principles Of Physical Biochemistry; Pearson/Prentice Hall, 2006.

(106) Kelly, S.; Price, N. Current Protein & Peptide Science 2000, 1, 349–384.

(107) Glasoe, P. K.; Long, F. A. The Journal of Physical Chemistry 1960, 64, 188–190.

(108) Kubelka, J. Photochemical & Photobiological Sciences 2009, 8, 499.